Solving the Catalogue Cross-Match Problem in the Galactic Plane

Thomas Jareth Wilson

Submitted by Thomas Jareth Wilson to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Physics, March, 2018.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signed: . . . . . .

Thomas J. Wilson

Date: . . . . . .

Abstract

1st Supervisor: Prof. Tim Naylor 2nd Supervisor: Dr Jennifer Hatchell

In this thesis the cross-matching problem is considered. One of the most fundamental processes in astrophysics, the cross-matching of two photometric catalogues is the assignment of an object in one catalogue and one object from the second catalogue as pairs, i.e., different detections of the same physical source in the sky. I present here new methods for considering such a problem, including the additional magnitude information available in photometric catalogues to break degeneracies between astrometric matches. I also generalise the Astrometric Uncertainty Function (AUF), usually assumed to be a Gaussian, to allow for the inclusion of systematic astrometric perturbations, such as those from blended sources of contamination.

The separations of sources in several widely used photometric catalogues with respect to the much more precise *Gaia* positions are considered. I find that the separations are described by a combination of a Gaussian distribution and a large non-Gaussian wing, and show that this is caused by flux contamination from blended stars not treated separately. At least one in three of the stars in the faint half of a given catalogue will suffer from flux contamination above the 1% level when the density of catalogue objects per point-spread function area is above approximately 0.005.

I then introduce a new method to use the additional photometric information from both catalogues in the process of accepting or rejecting counterparts, providing approximately a factor 10 improvement in Bayes' factor with its inclusion. The method uniquely combines photometric information from both catalogues while avoiding the use of prior astrophysical knowledge. Additionally, I formally describe the probability of two sources being the same astrometric object, allowing systematic effects of astrometric perturbation (by, e.g., contaminant objects) to be accounted for. I apply this method to two key match

cases, of two catalogues of similar wavelength coverage but differing dynamic ranges, and of two catalogues with approximately equal astrometric precision, discussing the importance of the inclusion of the magnitude information in each case.

Finally, the extension to the inclusion of perturbations due to faint contaminant stars in the AUFs of catalogues is combined with the improved cross-matching method for the specific case of the *Wide-field Infrared Survey Explorer* (*WISE*) catalogue. I describe the rigorous construction of the description of astrometric offsets due to faint stars, and then apply the method to *Gaia*-*WISE* matches in the Galactic plane. I analyse several test cases and discuss the photometric effects of the blended star contamination, showing that stars with significant astrometric perturbation are detectably photometrically compromised. I discuss the implications this has on derived parameters in several areas of astrophysics.

Contents

List of Figures

List of Tables

Declaration

I gratefully acknowledge support from an STFC Studentship and a CEMPS Ph.D. Studentship from the University of Exeter, without which none of this would have been possible. This thesis has made use of the SciPy (Jones, Oliphant, and Peterson, 2001), NumPy (van der Walt, Colbert, and Varoquaux, 2011), Matplotlib (Hunter, 2007), and F2PY (Peterson, 2009) Python modules, and NASA's Astrophysics Data System.

The work presented in Chapter 2 is taken from a paper published in the Monthly Notices of the Royal Astronomical Society by Tom J. Wilson and Tim Naylor, entitled "The Effect of Unresolved Contaminant Stars on the Cross-Matching of Photometric Catalogues" (MNRAS, 468, 2517). The work presented in Chapter 3 is taken from a paper published in the Monthly Notices of the Royal Astronomical Society by Tom J. Wilson and Tim Naylor, entitled "Improving Catalogue Matching By Supplementing Astrometry with Additional Photometric Information" (MNRAS, 473, 5570). The work presented in Chapter 4 is taken from a paper submitted to the Monthly Notices of the Royal Astronomical Society by Tom J. Wilson and Tim Naylor, entitled "A Contaminant-Free Catalogue of *Gaia* DR2-*WISE* Galactic Plane Matches: Including the Effects of Crowding in the Cross-Matching of Photometric Catalogues". The rest of the work presented in this thesis is my own unless otherwise stated.

This research has made use of the APASS database, located at the AAVSO web site. Funding for APASS has been provided by the Robert Martin Ayers Sciences Fund. I would also like to thank the team personally for their support and feedback during the early stages of this work. This work makes use of data obtained as part of the INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS, www.iphas.org) carried

out at the Isaac Newton Telescope (INT). The INT is operated on the island of La Palma by the Isaac Newton Group in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias. All IPHAS data are processed by the Cambridge Astronomical Survey Unit, at the Institute of Astronomy in Cambridge. The bandmerged DR2 catalogue was assembled at the Centre for Astrophysics Research, University of Hertfordshire, supported by STFC grant ST/J001333/1. This work makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This work makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (http://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, http://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This work is based [in part] on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA.

Acknowledgements

First, I have to thank Tim for all of his support through this PhD – both official and unofficial. This simply wouldn't have been possible without him, or any of the "aha!" moments that led to a better understanding of the problems that I encountered during these last few years, many of which are directly responsible for the final state of the work presented here.

My next thanks have to go to 407, for making office work seem slightly less (sometimes a bit too much less!) like work; to many a film or game night, to pub nights, to coffee time, to Sporcle, to AstroDnD, to Rainbow, to crosswords, to ridiculous conversation and to everyone who made those things what they were, for being there when I needed to not be working; and to all of my flatmates, past and present, for giving me a home full of great times.

Finally, I have to thank Hannah, without whom none of this serves much purpose, and who is the best thing to come out of my time in Exeter.

Tom J. Wilson

Exeter, U.K.

28th March 2018

Chapter 1

Introduction

The story so far: In the beginning, the universe was created. This made a lot of people very angry and has been widely regarded as a bad move.
— Douglas Adams, The Restaurant at the End of the Universe (1980)

1.1 Observing the Heavens

Since the dawn of time people have recorded images of the night sky above them. Some of the earliest clear archeoastronomical evidence available dates back to 300BC, from Namoratunga in north-west Kenya, where 19 pillars were found to be erected aligned with modern day constellations (Lynch and Robbins, 1978), a fascinating example of pre-historic documentation of the heavens. Further back in time still there are numerous examples of records of the night's sky: the Puyang Tomb, carbon-dated to 5300BC, suggested to contain carvings of constellations; the Nebra Sky disk, from Saxony-Anhalt, Germany, dated to circa 1600BC, depicting the Sun and Moon as well as a cluster of stars interpreted as being the Pleiades cluster; Stonehenge, a Neolithic ring of standing stones, perhaps used as a calendar for the motion of the heavens; and a depiction of a celestial diagram on the ceiling of the Tomb of Senemut from circa 1500BC. The best-known, and only surviving Wonder of the Ancient World, the Great Pyramids at Giza are hypothesised to have celestial significance as well. The alignment of the three pyramids, almost but not quite in alignment, has been suggested to map constellations visible above the Egyptian

horizon, with theories suggesting they align with the ancient positions of Orion's belt (and other, smaller pyramids aligning with Orion as a whole), or the Big Dipper, known to the Egyptians as "the Foreleg", represented by the "foreleg of ox" hieroglyph.

This thesis concerns the mapping of sources in the sky and the matching between different catalogues. Catalogues have a long history of their creation, updates to technology used to create them and the mathematical tools used to describe them, and the way astronomers used them. Before the issue of matching such documentations of the heavens can even be considered, they must be understood in their own right. Even before the invention of the telescope many astronomers still produced great lists of the objects visible in the sky. Perhaps the first was produced by the Greek astronomer Ptolemy in A.D. 150, although his original manuscript has long since been lost to the ages; it survives in the form of Ptolemaeus (1515), the Latin translations of Arabic documents in 1175 by Gerard of Cremona, themselves translations of Ptolemy's original Greek. His work lists 1,028 objects in 48 constellations – the 12 traditional zodiacal constellations plus 36 additional ones – although three of these objects are duplicated across two constellations. A further three objects are actually not stars at all; they consist of the Double Cluster h and χ Persei in Perseus, Praesepe in the constellation of Cancer, and the globular cluster ω Cen – all bright enough but dense enough to be visible to the naked eye yet not resolved as a more complex physical structure.

Ptolemy's work stood for almost 1,400 years until Tycho Brahe produced his "thousand-star" catalogue. His improvements to celestial chart making, as well as a sheer amount of hard work, resulted in a tenfold increase in the accuracy of the star positions. Originally circulated in 1598 in handwritten form, an abridged version of 777 of the most accurate positions was published posthumously, the year after his death in 1601 (Brahe, 1602). Johannes Kepler, his assistant at the time, continued the work he had started and finalised the publication of the full catalogue. Kepler and Brahe (1627) produced tables of such unprecedented celestial precision, aided by the new computational power of logarithms, accurate to within one arcminute – roughly the resolution of the

human eye – in most cases, that they were able to predict the transit of Mercury in 1631 and the transit of Venus the following year.

In the following century astronomers continued to produce increasingly accurate lists of observations. A significant amount of work on these observations was done at the Royal Observatory Greenwich by the Astronomers Royal. Created in 1675 by King Charles II after he founded the observatory, the Astronomers Royal were, until recently, the directors of the observatory. Its first Astronomer Royal was John Flamsteed, appointed by Charles II "forthwith to apply himself with the most exact care and diligence to the rectifying the tables of the motions of the heavens, and the places of the fixed stars, so as to find out the so-much desired longitude of places, for the perfecting the art of navigation". He spent almost forty years meticulously observing the sky above Great Britain and – after preliminary results were published by Isaac Newton and Edmond Halley without his permission – the Flamsteed (1725) catalogue was published by his wife, six years after his death. The first act of the third Astronomer Royal, the Reverend James Bradley, was to commission a new, eight-foot quadrant, improving upon the previous equipment of the Royal Observatory; the new quadrants could be read to within a second of arc. Bradley also spent much of his time in Greenwich cataloguing the sky, although politics interfered with the publication of his results. He worked often with Friedrich Bessel, who frequently offered astrophysical interpretations for Bradley's observations. The delays to the publication of his works meant that they too were published posthumously, in two volumes: the first a collection of 12 years of observations (1750-1762; Bradley, 1798), the second extending to include two years of observations made by his Astronomer Royal successor, Rev. Nathaniel Bliss, published seven years later (Bradley and Bliss, 1805).

Perhaps the most famous example of the documentation of the universe is that of Galileo Galilei (1610), and his recording of the moons of Jupiter over several nights in the winter of 1609-1610. It was his work on the telescope that enabled such detailed description of such a complex system, completely invisible to the naked eye. Over the centuries telescopes were improved and enlarged, allowing for the viewing of fainter

and fainter objects, far out of reach of viewing by the human eye under even the most ideal viewing conditions, offering yet more improvements over the catalogues created before them. However, the basic concept remained unchanged: no matter how large the magnification, it was the responsibility of the observer to record the images the telescope presented to them. This paradigm changed on July 17th, 1850, when William Cranch Bond and John Adams Whipple took a daguerreotype of the star Vega, kickstarting the era of astrophotography. These photographic plates vastly improved both the accuracy and efficiency of celestial observations, with Harvard Observatory producing the first sky survey between 1882 and 1886, reaching stars as faint 15% the brightness of the faintest possible star observable with human eyes, or 8th magnitude. The magnitude system dates back to the ancient Greeks and Hipparchus in 129BC. It had, however, been updated in 1856 by Norman Pogson, who proposed a logarithmic scale to the brightness of celestial sources, with a scaling factor of $\sqrt[5]{100} \simeq 2.512$, with the star Vega having a brightness of zeroth magnitude. At the same time, Bond realised that the size of a star on the plates went logarithmically with brightness (e.g., Ross, 1922), allowing for the assignment of relative brightnesses, or magnitudes, to the stars observed.

After the initial success of the Harvard Observatory, many other such observatories sprang up, such as that on Mount Wilson in California, USA. Much work was done using these new telescopes, examining the universe through – now much larger – telescopes. The rate of progression was so rapid that Bond's 15-inch Great Refractor of 1847 was overshadowed in 1917 by Mount Wilson's 100-inch Hooker telescope, almost a 50-fold increase in light collecting area. Critical to the research being done at these new observatories was comparisons between the results of observations taken at multiple observatories, ensuring that the interpretations from different data sources could be viewed on an equal footing (Seares, 1914). In 1887, the Astrographic Congress was held in Paris, with more than 20 observatories agreeing to participate in the creation of the Astrographic Chart. Over several decades across 22 sites, the Chart was slowly completed (Turner, 1912), with each observatory contributing hundreds of thousands of detections. The Astrographic chart

provided positions and brightnesses for the entire sky down to roughly 11th magnitude. As ambitious as this Chart was, proposed a mere decade after the first photographic plate observations were undertaken, its successor was even more ambitious. The Carte Du Ciel – literally the *Map of the Sky* – was to follow the Astrographic Chart, building upon its work but with detections of all stars as faint as 14th magnitude – a factor 16 fainter in flux. However, it would prove simply too big a challenge to complete, and the Carte was never finished. Each observatory keeps records of its own plates, and they are becoming an increasingly valuable source of historical brightness and position records (Geffert et al., 1996; Gavras et al., 2010), even being used in conjunction with the Hipparcos satellite (Dick et al., 1993).

However, in all of these surveys and throughout all of the work done until the first half of the 20th century, photographic plates were utilised on a plate-by-plate basis. The research carried out often only required one observation for its analysis, with many of our insights into the universe gleaned in these decades from relatively simple analysis, at least by modern standards. In cases where two plates needed to be combined, there was likely only one source of interest, the brightest sources in the sky being sparsely arranged, and as such it was trivial to combine the information required. Thus began the story of the catalogue cross-match, albeit in a vastly different form. The origins are simple and intuitive enough: for a given plate the region of the sky in question is known, and thus a different plate of the same area of sky can be located. Then it is a simple matter of finding the one, or perhaps two, stars detected and noting their position – or positions – relative to the plate, and noting the similar positions between detections. This basic idea never changes; the method is simply updated to account for changes throughout the remaining 60 years of its history.

1.2 Finding the Counterpart to Sco X-1, and Other Difficult Identifications

The photographic plate sky was so sparse that there were no problems confirming the identity of a star seen on a new plate exposure in a plate taken by another astronomer at a different observatory. While not especially precise by today's standards, the location of stars was known sufficiently to be able to distinguish between sources in the great lists created in the late 1800s and early 1900s. It was not until the second half of the 20th century that the astronomical community had cause to think any differently about its source identification methods, with the advent of observations in both X-rays and at radio wavelengths. These detections were subject to extremely low levels of positional confidence, suffering three or four orders of magnitude less precision than that of the photographic plates. Perhaps the most famous example of these early detections is that of Sco X-1, the first X-ray detection outside of our solar system. Its initial detection was constrained to only slightly better than a tenth of a percent of the entire sky – an area 100 times larger than the full moon. A new strategy for the identification of the sources of these emissions in the optical was required.

Bruno Rossi proposed the existence of celestial soft X-ray sources in 1958. He urged the United States Air Force to explore the lunar surface in the years leading up to the launch of astronauts to the Moon. The proposed mission would also be able to search for additional sources of X-ray emission, and thus in June 1962 a team headed by Riccardo Giacconi launched an Aerobee 150 rocket, finding significant X-ray emission off centre of the moon at 16 hours and 15 minutes Right Ascension, -15.2° Declination (Giacconi et al., 1962).

The detection was initially confused with the Galactic centre due to the lack of collimation on the detectors used aboard the rocket. The Moon, as the primary target, was deemed too large to require such additional measures. However, with subsequent experiments (June 1963, Gursky et al., 1963; and August 1964, Giacconi, Gursky, and Waters,

1964) it was determined that the source originated from the constellation of Scorpius, and thus the source was dubbed Sco X-1 – the first X-ray source in Scorpius. The nature of the X-ray emission was uncertain, with initial hypotheses ranging from synchrotron radiation, to X-rays from the radio centre of the Galaxy, to isotropic extragalactic flux.

To better understand the source and confirm the origin of these X-rays, it was determined that the source must be identified in the optical. Unfortunately, while the typical photographic plate measurement was good to within 5 arcseconds, the position of Sco X-1 was initially only known to within 5 degrees, constrained with repeated experiments to an uncertainty of 30 arcminutes. Four years after its initial discovery, the source of the X-ray emission was finally confirmed by Sandage et al. (1966). In all, it took 11 nights over two months, using four telescopes and three observatories to obtain the data necessary to draw any meaningful conclusions. A bright ($V \simeq 13^{\text{th}}$ magnitude) object was detected approximately one arcminute from one of the possible locations of the X-ray emission, well within the half degree error box. It was found to be variable, and assigned the designation V818 Scorpii (Kukarkin et al., 1968), and subsequently identified as a neutron star (Shklovsky, 1967). Today, the neutron star is known to be one half of an X-ray binary (e.g., Steeghs and Casares, 2002). However, it in the late 1960s, it took several years and numerous attempts by teams of researchers to identify the object found in one survey in the same region of sky observed as part of a different survey. This difficulty was further compounded by the fact that you had to know what to look for in the follow-up spectra when using them to identify the correct counterpart; you had to know what you expected the answer to be before you asked the question.

The story of X-ray emission counterpart identification continues into the early 1970s. With a rapidly expanding set of potential X-ray sources discovered, the optical counterparts to many new astrophysical objects were required to better understand the ensemble of emitters. By 1970, six optical counterparts, as well as a single radio counterpart to X-ray sources had been identified: Sco X-1; the supernova remnants of the Crab Nebula, Casseiopeia A, Tycho's Nova, Cygnus Loop; and the Crab Pulsar (Kunkel

et al., 1970). However, progress on identification was slow, with many searches returning no clear candidate for the emission (e.g., Kunkel et al., 1970; Peterson, 1972). X-ray candidate confirmation remained a difficult process, taking many nights of spectroscopic and photographic observations (Webster et al., 1972).

This story is mirrored in early radio astronomy. The first extrasolar radio waves were discovered, purely by chance, by Jansky (1933). Investigating interference in short wave voice transmissions, Jansky realised he was recording a repeating signal on a cycle of 23 hours and 56 minutes, the sidereal day. He eventually concluded that the source must be interstellar gas and dust in the direction of the constellation of Sagittarius; this source is now known to be Sagittarius A, emission from particles orbiting the black hole at the centre of the Milky Way. A decade later, in 1942, during the Second World War, both Hey (1946) and Southworth (1945) detected radio waves from the Sun, although, as both were bound to wartime secrecy regarding the capability of radar, Reber (1944) was the first to publish such findings.

During the 1950s the first surveys were conducted, with the release of the 2C (Shakeshaft et al., 1955) and 3C (Edge et al., 1959) radio surveys using the Cambridge Interferometer. Radio detections suffered similar levels of sky position confusion, with 15 arcminute uncertainties not being uncommon (e.g., Dewhirst, 1959). The optical identification of the first quasi-stellar radio sources ("Quasars", first used May 1964), expected to follow much the same tale as that of Sco X-1, was much less challenging. The first Quasar – although not identified as such at the time – was initially discovered by its radio emission in the 3C survey, being designated 3C 273, the 273rd detection by the survey, with roughly 5 arcminute precision. The second Quasar to be identified, 3C 48, had similar precision in the 3C survey results. It was 3C 48 that had its optical counterpart identified first, however, with the utilisation of two 90-foot antennas creating an interferometer that allowed for the localisation of 3C 48 to better than 10 arcseconds (Matthews and Sandage, 1963), much reducing the source follow-up challenge facing a multi-arcminute precision detection. Serendipitously, it was determined that 3C 273

was to be occulted by the Moon, and thus in the same year its position was measured to approximately one arcsecond (Hazard, Mackey, and Shimmins, 1963), with its optical counterpart being found by Schmidt (1963) with relative ease once its position had been found to such high precision.

Entirely analogous to the searches for X-ray emission, the optical counterparts to radio sources typically required extensive follow-up. The 3C catalogue of sources, good to a precision of a few arcminutes, required follow-up with radio interferometers to allow for a more precise radio position to be determined. Alternatively, spectroscopic analysis was usually necessary to identify the specific source responsible for the radio emission amongst the multitude of potential sources in the tens of square arcminutes of sky around the radio emission, even through to more modern times. Even as recent as 2005, 200 square arcminutes of sky was searched around each source in the HIPASS catalogue for spectroscopic targets (Doyle et al., 2005). These early days of multi-wavelength source identification therefore required significant time and energy to be expended to robustly confirm or reject an optical source as the X-ray or radio wavelength counterpart. The rapid strides made in understanding these new and exotic astrophysical sources are all the more incredible when put into the context of the tools the astronomers had at their disposal.

1.3 The X-ray Satellites and the Brightest Star in the Sky

For the most part, the difficulties faced in searching for the same astrophysical object in two vastly different surveys were simply a product of the technological limitations of the time, with both X-ray and radio astronomy in their infancy, and follow-up observations being time consuming and arduous. It wasn't until the end of the 1970s, with the successful launch of HEAO-2, later re-named the *Einstein* Observatory, in November 1978 that an improvement in X-ray astronomy was achieved. During its three year mission it offered two orders of magnitude greater astrometric precision than any other X-ray observations before it. Following its success came *EXOSAT* in May 1983 (Taylor et al., 1982) and *ROSAT* (Aschenbach et al., 1981), eventually launched in June 1990, offering better than

5 arcsecond precision during its eight year life. This was, in turn, followed by the *Chandra X-ray Observatory* (Weisskopf et al., 2002) and the *XMM-Newton* telescope (Jansen et al., 2001), both still in operation nearly 18 years after their launches in 1999.

Together these missions revolutionised the way X-ray astronomy was undertaken; but they also completely changed the approach to the counterpart search. The reduction in uncertainty of position, by as much as a factor of 100, led to a corresponding reduction in sky area for potential counterpart search of upwards of 10,000. With repeated observations, now possible due to the extended life of this new generation of X-ray missions, the uncertainty on a source's position could reach as low as 2 arcseconds, as in the case of the determination of the optical counterpart to LMC X-1 with *ROSAT* (Cowley et al., 1995). This meant that there would typically only be 2-4 potential counterparts for follow-up, with one object much more likely to be the source of the X-rays (e.g., Feigelson and Kriss, 1989). Where multiple counterparts existed, typical observations involved the "extremely naive" strategy of following up observations in descending brightness order (Stocke et al., 1983). When choosing how to distribute what time observers had been allocated on optical telescopes, they often chose to prioritise the brightest objects in their error boxes, as when Buckley, Tuohy, and Remillard (1985) initially took a cut of stars of tenth magnitude when following up X-ray sources, or Mason et al. (1995) assigning a $V = 14$ cut to follow-up spectroscopy using the *Isaac Newton* Telescope, La Palma.

This strategy was not without scientific merit, as the longer wavelength – most often optical, but yet longer wavelengths, into the infrared, were utilised in situations of heavy obscuration – counterparts were often found to be very bright. The infrared (IR) counterpart to the Galactic bulge low-mass X-ray binary GX 13+1 was identified as having a K-band magnitude of $\simeq 11.5$ and the IR counterpart to GX 5-1 has $K \simeq 13.5$ (Garcia et al., 1992; Naylor, Charles, and Longmore, 1991). Indeed, the *ROSAT* team, following the experience gained from the *Einstein* Observatory, recommend the X-ray-to-visual flux ratio (f_x/f_v) as a diagnostic for breaking degeneracies between multiple potential X-ray counterparts (Danziger et al., 1990).

While the earlier missions still required spectroscopic follow-up to confirm the large-scale population trends (Bouvier and Appenzeller, 1992), eventually, with the large-scale surveys of *ROSAT*, *Chandra*, and *XMM-Newton*, the population of X-ray sources was sufficient that the statistical samples of astrophysical sources could be analysed. Brusa et al. (2007) used a sample of approximately 700 *XMM-Newton* sources to create a broad distribution of optical counterpart magnitudes. Using almost 19000 *ROSAT* sources, Haakonsen and Rutledge (2009) created an empirical distribution $N(< m)$, the number of IR sources brighter than the potential counterpart magnitude, assigning brighter sources a higher weighting than faint objects. This method was summarised succinctly by Fotopoulou et al. (2016): "...X-ray sources are rare events; bright optical sources are also rare events, so the observation of an X-ray source and a bright optical source in the same region of the sky is considered a non-random event".

1.4 Winning the Error Box vs Depth Race

Stepping away from the more troublesome wavelength ranges of the electromagnetic spectrum – avoiding X-ray and radio detections – the problems surrounding source cross-catalogue identification are significantly lessened. Observations taken in more manageable parts of the electromagnetic spectrum, primarily in the optical wavelengths, continued to be much more straightforward into the 20th century. During the decades of the 1970s and 80s one of the primary advancements in astronomy was the invention of the Charge-Couple Device (CCD; Boyle and Smith, 1970). By 1976 these new devices had already been used to observe astrophysical sources, such as the planet Uranus (Smith, 1976). While this technology focussed on the optical wavelengths, the infrared spectrum was coming into its own as an astronomical pursuit as well, with stellar photometry in the near-IR being developed in the 60s (e.g., Johnson, 1962, defining the J, K, L, and M passbands). IR array detectors followed the optical CCD nine years later (Forrest et al., 1985). In addition, while the 200-inch Hale reflector on Mount Palomar was the largest telescope in the world for 27 years, the 1970s and 80s saw the first light of numerous 4- and

5-metre class telescopes, such as the Russian-built Large Altazimuth Telescope, 1975; the Multiple Mirror Telescope built in 1978 in Arizona, USA; the United Kingdom Infra-Red Telescope (UKIRT) and Canada-France-Hawaii Telescope, built on Mauna Kea, Hawaii in 1979; and the *William Herschel* Telescope built on La Palma in the Canary Islands in 1987. These large telescopes, continuing the trend started by the Mount Wilson and Mount Palomar Observatories of building telescopes in places of much more favourable atmospheric conditions than the North-East of the USA, combined with the improvements offered by CCDs over the photographic plates of old, vastly increased the faint limiting magnitude of potential observations while increasing the accuracy of the positions of sources in individual image exposures.

Engineering feats improved the precision of *relative* astrometry, the positions of detections of a given exposure relative to its other detections. However, it was the conceptual improvements to the celestial reference frame used to translate the positions of stars on a plate, and later CCD, to celestial coordinates that improved the *absolute* astrometry of photometric catalogues. Optical catalogues typically only relied upon relative astrometry to identify cross-catalogue pairs, while the sparsity of X-ray detections forced any positions to be described in absolute astrometric coordinates. The 1980s saw work on increasing the precision of the absolute astrometry of the network of bright guide stars used to construct the transformations from cartesian to celestial coordinates. This started with the Carlsberg Automatic Transit Circle (Morrison and Gibbs, 1986), with Hipparcos further increasing the number of stars available for reconstructing the absolute astrometry of a source. This increase in the number of "guide stars" available for astrometric reconstruction was mirrored in field-of-view (FOV) increases of the telescopes of the 1970s and 1980s, compared to the previous generation, which allowed for a more robust astrometric solution to be calculated for a single exposure. While the relatively new CCD technology meant that the new detectors were smaller than the photographic plate that came before, these newer telescopes had smaller f-numbers than previous telescopes; thus despite its much smaller detector, UKIRT had a larger FOV than the 200-inch Hale

reflector on Mount Palomar. However, even these improvements were themselves subject to systematics that had to be understood when combining the absolute astrometry of catalogues of differing wavelengths. The "anchor" points used to define the references frames for telescopic pointings, by necessity of the varying emission of celestial sources, differ across various wavelength regimes. At optical and IR wavelengths, external galaxies can be used, providing absolute astrometric coordinates with high precision due to their abundance; however, at radio wavelengths the radio-loud jets from the black holes at the centre of these galaxies must be used, which can lead to positional offsets if these jets are spatially separated from their host galaxy. The International Celestial Reference Frame (ICRF) is defined using 212 extra-galactic sources, most of them Quasars, which can also be used to anchor relative astrometry to absolute astrometric coordinates in the X-ray regime, crucial for improving the counterpart matching of sparse X-ray sources.

The transformation from relative astrometry and source coordinates in a single image to absolute astrometry, allowing for the comparison between sources across the entire sky, comes at a cost, however. Until this new generation of telescopes was built, the individual uncertainties on a position had been sufficiently large than they could be considered the end of the story – the certainty with which a source can be located on a photographic plate was the dominant term. However, with these technological improvements the brightest stars could be pinpointed with increasing precision. We therefore must now, perhaps for the first time, consider the effects of the astrometric "plate solution" – although in the 1980s observations were not frequently done with photographic plates, the name has stuck to this day – on the positions of sources. Until now, when discussing the astrometric uncertainty of a source, I have not distinguished between the uncertainty in pixel space and the uncertainty in coordinate space, as it is often a simple application of a "plate scale" – some number of arcseconds to a pixel side – and if a source has an uncertainty of two pixels it has an astrometric uncertainty of twice the plate scale. However, when calculating the transformation from pixel to astrometric coordinates there is some residual uncertainty, caused by a combination of the separations between all sources used to construct the

solution in both the image and the guide star catalogue. This uncertainty must be included in the positions derived for all individual sources in the image, even if they were not used in the solution construction. Therefore a star truly has a positional uncertainty that is a combination of its *statistical* uncertainty – the same uncertainty I have been implicitly discussing throughout this chapter, the uncertainty of the individual centroid of a source on the CCD – and its *systematic* uncertainty – the uncertainty of the global solution. For the brightest sources the statistical uncertainty can be very small, such as in cases where its PSF is characterised well and the pixel scale is small enough to sample the PSF structure, with centroiding possible below a tenth of a pixel. These bright sources therefore have a dominant systematic uncertainty, and thus their positions cannot be determined to below the level to which the global astrometric coordinate system has been applied to the entire image.

The improvements to both the statistical and systematic sides of the centroiding methods led to sub-arcsecond positional uncertainties and typical observations reaching 16th magnitude in the infrared (Leggett and Hawkins, 1989) – with the rare exception of much deeper surveys, such as the 22 hour integration time employed by Cowie et al. (1990) to reach $K = 21$. The typical spacing between stars, relative to their now seemingly pinpoint positions, was larger than ever. Not since the 1880s could a star in one image be so easily identified as the same star in an opposing survey's images, but now this could be achieved with stars 8 magnitudes, or a factor of 1600, fainter.

Perhaps the only exception to this trend of precision outpacing sensitivity is that of the longer wavelength – mid- to far-IR – space-based missions. The first infrared space telescope, the *Infrared Astronomical Satellite* (*IRAS*) began operation in 1984 (Neugebauer et al., 1984). The next telescope, the Infrared Space Observatory (ISO; Kessler et al., 1996), launched a decade later, offered over an order of magnitude spatial resolution improvement and three orders of magnitude increase in sensitivity. However, after ISO, the infrared space observatories are limited to similar spatial resolutions (AKARI, Murakami et al., 2007, launched in 2006) and sensitivities, as with the *Wide-field Infrared Survey*

Explorer (*WISE*; Wright et al., 2010), launched in 2009. The *Spitzer Space Telescope* (Werner et al., 2004), launched in 2003, offered a three-fold improvement in spatial resolution over these other infrared space missions. This improvement is due mostly to its shorter wavelength coverage – 3.6-8μm, compared to *IRAS* operating at 12, 25, 60, and 100 μm or ISO operating out to 240 μm – and size – having a 0.85 metre diameter, significantly larger than the 0.4 metre diameter of *WISE* or the 0.57 metre diameter of *IRAS*. This improvement in precision, however, was offset by a five-fold increase in sensitivity, increasing the number of potentially overlapping sources detected. Although the race between precision and flux limit is not quite as clearly won, space-based infrared telescopes, much like their ground-based counterparts, still offered increasingly precise astrometry in the 1990s and early 2000s.

It therefore seemed that, so long as the science being undertaken didn't require detections of the chosen source in X-ray or radio wavelengths, the building of a comprehensive picture of an astrophysical source across multiple filter images or across multiple telescopes would remain, for the most part, as straightforward as it has ever been. The merging of two datasets, each containing a number of stars with photometric magnitudes, astrometric positions, and their related uncertainties has become a fundamental process in many aspects of astrophysics. Broadband photometric measurements are crucial to gaining an understanding of a whole host of phenomena, from stellar physics to extragalactic luminosity functions.

The simplest matching method only utilises the knowledge of the stars' positions and use a nearest neighbour approach with a maximum cutoff distance when matching stars between two catalogues. In the late 19th century this method was perhaps more intuitive, with one or two detections per plate, but could be easily extended to wider field observations with increasing numbers of detections. Within the critical separation, two stars in two catalogues whose closest star in the other catalogue is each other will be assigned as a match, without consideration of either catalogue in a wider context, just considering each match on a pair-by-pair basis in isolation. I shall refer to this as "nearest

neighbour matching" throughout this thesis. It is often also referred to as "proximity matching" for obvious linguistic reasons, but this term is, in uncommon cases, used interchangeably with a different method, and the conflicting terminology can be avoided with the more robust language.

For some years this method has been used with relative success, owing for the most part to its simplicity. It was used by Miller, Margon, and Burton (1993) to identify the IR counterpart to GX 340+0, 22 years after its discovery in 1971 as an X-ray source. With the position being known previously to within an arcsecond, the largest source of uncertainty in the association was the K-band detection, known with 1.5 arcsecond precision. The search radii of such nearest neighbour matches can vary considerably, from very tight matches (e.g., 1 arcsecond, Dong et al., 2011; 3 arcseconds, Cutri et al., 2012; 6 arcseconds, Theissen, West, and Dhital, 2016) to larger radii (e.g., 16.5 arcseconds, Kellogg et al., 2015; 1 arcminute, Mocanu et al., 2013).

The use of telescopes to construct catalogues of objects on entire sky scales became increasingly common towards the turn of the 21st century, perhaps most famously with the Two Micron All Sky Survey (2MASS; Skrutskie et al., 2006), running from 1997 to 2001. The datasets provided by these large surveys contain the derived positions and brightnesses of vastly more data than could ever have been imagined at the beginning of the 20th century. Such detections, of differing wavelength coverage, resolution, dynamic range, etc., are still used together to maximise scientific potential. This merging process is the "cross-matching" of the catalogues, through which detections across several surveys corresponding to the same astrophysical source are identified and combined, much as they have been since the dawn of astrophysics. No matter how much data was generated, it still seemed that the race was won for a lot of cases; error boxes shrunk at a rate quicker than stars were added to the skies, with few exceptions. So long as the catalogues constructed probe the same sources – both astrometrically and philosophically – on the sky then the creation of the composite dataset was an open-and-shut case, without the complexities that had gone before.

1.5 Probability-Based Catalogue Matching

Unfortunately, the differences between datasets of different surveys introduce difficulties when constructing a merged dataset, which must be accounted for in order to not introduce systematic effects. Additionally, the naive nearest neighbour matching scheme has several limitations. Its primary issue is that it does not consider the possibility that the closest object is not the correct object. Furthermore, despite the fact that there might be an object in the second catalogue within the critical radius, the source in question could have properties that would place its detection outside of the dynamic range of the second catalogue. This is becoming increasingly a problem in more recent years with the latest, faintest surveys conducted to date. Observations using the Dark Energy Camera on the Blanco telescope in Chile reach 5σ detections as faint as 25th magnitude, but saturate corresponding fainter, at around 15th magnitude. If this dataset were cross-matched with the AAVSO Photometric All Sky Survey (APASS; Henden and Munari, 2014), with typical faintness limits of only 16th magnitude, then there would be little overlap in the good quality data from these two catalogues. In fact, it might be reasonable to use *both* surveys for scientific purposes, now having detections of optical sources in a 20 magnitude dynamic range, compared with the 10 magnitudes offered by either individual dataset. It might also be desirable to use our older, less precise datasets. These may offer time-series observations, legacy results (such as the brightness of sources pre-outburst or supernovae), or simply still be the most useful data available to the research in question.

Given these issues facing these large-scale surveys, it is perhaps salient to conceive of a quantifiable way to rate or otherwise rank the pairing associations between two catalogues. For cases where one measurement has high uncertainty, leading to multiple potential counterparts, this can give a relative score to each potential counterpart, without the time-consuming task of spectroscopic follow-up. It is also valuable for those high-precision surveys, quantifying the likelihood that the closest object to a given source in the second dataset is spurious, or the real source is missing (due to either bad quality data, or its intrinsic brightness being below the sensitivity of the survey, for example) and the

matched object is unrelated.

This scheme of quantitatively ranking pair associations has its origins in the late 70s with the characterisation of the optical counterparts to radio sources (de Ruiter, Willis, and Arp, 1977) and the 80s with the identification of *IRAS* sources (Wolstencroft et al., 1986), continuing the likelihood ratio (LR) method first proposed by Richter (1975). The method considered the balance between the chance of a true counterpart being found at some separation from its corresponding second survey detection and the chance that an uncorrelated, random second source would appear at that same distance. Varying functions have been used for the two halves of the ratio, with the nearest neighbour chance sometimes being parameterised as a Poissonian distribution or a simple "background" source density, for example. Rarely, the random nearest neighbour likelihood has been used on its own to quantify the chance of source association (Webb et al., 2003); this occurs much less frequently in the literature than the use of the ratio, however.

Probabilistic catalogue matching is commonly accepted to have first been quantified properly when Sutherland and Saunders (1992) discussed the problems with matching optical data to non-optical sources. The first to lay out the method in detail, their teachings are cited to this day. Unfortunately, this excellent discourse on matching probabilities was not immediately recognised; it took over three years for a citation to the paper to appear in the literature, and almost 20 years before the astrophysical community acknowledged the importance of this 26-year-old work properly, receiving over 70% of its citations in the last eight years. To overcome incorrect matches, Sutherland and Saunders (1992) defined the reliability of a source. They used knowledge of the source's "type" to identify optical counterparts to *IRAS* and radio galaxies, and overcome any faint object being assigned as a counterpart by nearest neighbour matching. This use of "typing" is advantageous when considering two very different catalogues. In the case of radio detections, almost all of them are external galaxies (e.g., the 3C catalogue; see Section 1.2 for more discussion), which allowed for a binary star-galaxy separation, each with very different optical properties. They also extended the LR method to the reliability with the inclusion of all competing

hypotheses. The LR is simply the balance of the pairing of two sources, one in each catalogue, to the non-pairing of those same two sources. It can take any given value, with values greater than one indicating the pairing is more favourable than non-pairing, values smaller than unity indicating the null hypothesis to be the correct one, and values on the order of unity suggesting no conclusions can easily be made. Including all hypotheses allows for the normalisation of the ensemble of likelihoods, giving true probabilities. Critically, however, the reliability can include the relative likelihood of *competing* stars being the counterpart to the chosen detection, which the LR simply cannot do; I discuss this in more detail in Chapter 3.

A thread in the literature (e.g., Rutledge et al., 2000, Fleuren et al., 2012) uses the reliability, the extension Sutherland and Saunders (1992) made to the LR method, to quantify catalogue matches, supplementing astrometric knowledge with magnitude information available to create one-directional relationships between different types of object and their brightnesses. For example, Naylor, Broos, and Feigelson (2013) map X-ray sources onto IR detections, using the magnitudes in the IR catalogue but not those in the X-ray data, mirroring and extending the concepts used by Brusa et al. (2007) to assign weighting to potential matches of a range of brightnesses. Their formalism for the matching procedure mirrors that of the X-ray counterpart identification discussion of the late 80s and 90s, showing mathematically that, indeed, X-ray sources are brighter in the infrared than the typical source.

Another thread follows asymmetrical matching using solely the likelihood ratio of counterpart pairs (e.g., Mann et al., 1997, Brusa et al., 2005). As previously mentioned, the likelihood ratio between two stars from different catalogues is independent of the close presence of a second object in one of the catalogues, and is therefore a suboptimal solution in cases of high source density. Where the chances of multiple sources being positionally close to a given object is high the assumption that the distances between stars are significantly greater than the matching radius holds in neither catalogue. All competing hypotheses must therefore be considered jointly if any conclusion about the likelihood of

an individual match is to be drawn, which may include the chance that multiple stars from either catalogue are potential matches to more than one star from the opposing catalogue. Naylor, Broos, and Feigelson (2013) include the explicit probability of a non-pairing of the X-ray source to any of the IR detections when considering such asymmetric multiplicity, as it is possible (such as in cases of differing dynamic ranges of the two surveys) that no source detected within the given error box of an object is its true counterpart.

Through most of the work undertaken in the literature quantifying the counterpart pairing likelihood, the addition of the distribution of brightnesses of only one catalogue is typically used. Naylor, Broos, and Feigelson (2013) and Brusa et al. (2007) consider the magnitude distribution of IR and optical sources when matching X-ray detections, for example. In neither case does the flux of the X-ray source factor into the formalism; it either has X-ray flux, or it does not. However, Budavári and Szalay (2008) symmetrised the procedure for the first time, considering magnitudes in both catalogues in question as equals to one another. They give an example of fitting the optical Sloan Digital Sky Survey (SDSS) to UV data from *GALEX*. While previous methods would, again, have considered a binary detection/non-detection for the UV fluxes, their formalism allows for the relative brightness in both the optical and the UV passbands to influence the confidence with which they assign their catalogue pairings. However, they used astrophysical information to do so, fitting theoretical spectral energy distributions (SEDs) to each hypothetical match. This fitting then leads to a merged catalogue that is dependent on the assumptions made about the theoretical models. This off-shoot in the catalogue matching problem has its own thread running through the literature as well, with the application of models to distinguish good matches from a host of degenerate astrometric potential matches. Marquez, Budavári, and Sarro (2014) use theoretical SEDs to fit optical and near-IR data for COSMOS galaxies, for example. While most of these methods focus on the matching of catalogues in the optical, IR or X-ray wavelengths, there are examples of matching in other wavelengths in the literature. These include Line et al. (2017) at radio wavelength and Pineau et al. (2017) more generally across catalogues with relatively precise astrometry.

These three methods – the original likelihood ratio method, the extension to the reliability by Sutherland and Saunders (1992), and the catalogue photometry symmetrisation offered by Budavári and Szalay (2008) – form the core of the literature using probability-based catalogue matching for scientific purposes in the last few decades. These each offer unique advantages and disadvantages, and can be better or worse suited to specific analyses, depending on the individual science case each user requires.

The LR method is the most intuitive, simplest mathematically, and least computationally expensive. However, it trades these bonuses off with its requirement of sparse datasets, and suffers from high false match rates, with increased false positive rate without the inclusion of photometric information (e.g., Wolstencroft et al., 1986) or in crowded fields where pair associations are less certain than can be assumed on an individual basis. In such a case two potential counterparts with equally high likelihood ratios would match individually but return a slightly less than 50% probability when including all possible hypotheses. In contrast, the reliability therefore decreases the false positive rate by allowing for all potential hypotheses of a given catalogue to be considered for a source in the opposing catalogue, although still with the requirement that the second catalogue be sufficiently sparse, for a slight increase in computational and mathematical complexity. Finally, the inclusion of theoretical photometric models allows for the generalisation of the two catalogues, allowing for the inclusion of the photometric information provided by both catalogues, improving the ability to distinguish between true and false counterparts, decreasing the false positive rate. This improvement is, again, traded off by an increase in computational complexity, as well as a slight potential increase in both false positive and negative match rates, based on poor assumptions made when generating the theoretical astrophysical models used to analyse the photometric probabilities. This method also requires knowledge of the astrophysical sources undergoing cross-match a priori, requiring initial effort to generate the theoretical models that produce the theoretical photometric information to achieve sensible cross-matches. In addition, all three methods suffer from some intrinsic level of false negative matches due to their non-treatment of systematic

causes of astrometric separation (e.g., proper motions), which will decrease the astrometric probability of those affected sources. Thus, as I discuss further in Section 3.1, there is a critical region of the matching algorithm parameter space currently not explored, which this thesis focusses on: a method with extremely low false positive and negative match rates, makes as few assumptions about its datasets (such as sparsity) as possible, and requires as few astrophysical assumptions as possible. This is, however, traded off by the necessity of increased mathematical and computational complexity; I discuss the mathematical framework in Chapter 3 and the computational costs in Chapter 5. This method will become increasingly necessary in the next decade with the next generation of telescopes and very faint photometric surveys, which I put into context in Section 5.1.

Despite the parallel development of the differing approaches to the quantifying of cross-match pairings, the commonality between all methods is the requirement for the description of the separation of two detections of the same source. Sources are defined by their detected sky position, as well as a corresponding uncertainty in this measurement. These methods fold in this extra information about the astrometric precision of such astrophysical detections, and the more complex formalism allows for the inclusion of the certainty to which the detections' positions are known. This can lead to the possibility that an object with a larger absolute separation from another source can have a smaller normalised sky offset – the ratio of its sky separation to the uncertainty in its position – than one that is detected closer to the source. To achieve these improvements requires the creation of probability density functions (PDFs) that describe the likelihood of stars being related – or not – based on their respective astrometric precisions and sky separation. This improves upon the static cutoff radius of the nearest neighbour match by taking into account the relative astrometric precision of the catalogues.

While the majority of the literature on the matching problem has been driven by astronomers for purely research-driven needs, there is, again, a separate drive in parallel from a computational perspective. These new, all-sky surveys result in datasets that number billions of entries, stretching computing power to its limit. It is therefore also necessary to

develop complex algorithms for efficient and timely identification of these sources. Ogle et al. (2015) discuss such a method in the context of the NASA Extragalactic Database, applying a rule-based matching algorithm to the problem to allow for matching that can be scaled to multiple catalogues across the entire sky in a tractable fashion. There has been considerable interest in this problem from a pure mathematical standpoint, reaching such a level as to merit a topic review, with Budavári and Loredo (2015) summarising the state of the statistical record linking problem – the more general case of the cross-matching of photometric datasets – over the past few decades, focussing particularly on the hierarchical Bayesian inference model.

It is therefore in recent decades that the pairing of detections of astrophysical sources in different surveys can be undertaken on a large, all-sky scale. The quantitative formalism assigns weight and relative scaling to the methods previously used, perhaps intuitively, in the latter half of the 20th century. This allows for the buildup of ensemble statistical descriptions of a variety heavenly sources – such as stars and galaxies – and the discerning of true and false counterpart assignments.

1.6 The Astronomical Error Function

While it is only in the last few decades that the probabilistic formalisms for the cross-identification of potential sources in the sky have been created, the history of the underlying descriptions is much longer. The most commonly-used PDF to describe the relative chance of association between two detections given the information about their positions and respective uncertainties is the Gaussian. For most of the 19th century, the function was referred to as the "astronomical error function", given its key usage in astronomy with many astronomers contributing to the foundations of the branch of mathematics it spawned. It is perhaps responsible the beginning of modern astrometry, with Bessel and Bradley (1818) using the newly developed theories to publish the most accurate positions of 3,222 stars measured to date, with Bessel's works remaining second to none for many years. He was the first to identify Sirius as a double-star system (Bessel, 1844), correctly

ascribing inconsistencies in the motion of the star over 90 years of observations to a hidden companion, following the discovery of its proper motion by Halley (1717) through comparison to Ptolemy's positions some 1,800 years earlier. However, for most of its history the astronomical error function was not used on star positions at all. The main focus of those working on the properties of this new function were concerned with the positions of the planets. Bessel's work, such a revolution in source precision as it was, led to the discovery of Neptune in 1846, following his research into the discrepancies with the orbit of Uranus undertaken in 1840 (Bessel, 1848).

The astronomical error function began, albeit in a vastly different form, with de Moivre (1733), when he wrote a private paper to some friends discussing the approximation of the sum of the terms in a Binomial expansion. In this paper, over seven pages, he concludes that if "...n = 3600, hence $^1/_2$n will be = 1800, and $^1/_2\sqrt{n}$ 30, then the Probability of the Event's neither appearing oftner than 1830 times, nor more rarely than 1770, will be 0.682688". However, it wasn't until Gauss (1809) that the mathematical groundwork was truly laid for the function, with the treatise he laid out so influential that the function bears his name to this day. In his essay he builds upon work on linear solutions done by the marquis de Laplace (1774). The work being undertaken still has nothing to do with the problem of catalogue matching, or indeed stars at all, as Gauss considered the orbits of heavenly bodies around the Sun. During his essay he developed the necessary mathematical tools, language and syntax to quantify their motions. He summarised thusly: "...If, for example, the measures of precision of the observations... have been found... the most probable system of values... will be that in which... the sum of the squares of the differences between the actually observed and computed values multiplied by numbers that measure the degree of precision, is minimum".

Laplace furthered the development of the work on this in his essays on the theory of probability (Laplace, 1820). In later decades, cementing the Gaussian as the astronomical error function, astronomers such as the Belgian astronomer Adolphe Quetelet and the Englishman Sir John Herschel – son of William Herschel and nephew of Caroline Herschel

– continued to research the applications of the function. In an address to the Edinburgh review, Herschel (1857) detailed the work Quetelet contributed over his career. He eloquently discussed the philosophical thinking behind the necessity of describing the uncertainty with which one can know the position of a star with a Gaussian distribution. His initial line of reasoning discussed "a ball dropped from a given height", describing the "probability of successively committing any given system of errors" and concluding that "the product of their separate probabilities must be expressed by the same exponential function of the sum of their squares however numerous, and is, therefore a maximum when that sum is a minimum". He links the discussion to the astronomical community at large, supposing "the rifle replaced by a telescope duly mounted... and we have the case of all direct astronomical observation where the place of a heavenly body is the thing to be determined".

It can therefore be shown that the spatial probability distribution associated with this type of problem is described by a Gaussian. These PDFs change based on the assumptions made about their form. The naive assumption is usually made that the astrometic uncertainty functions (AUFs), as I shall refer to them – as opposed to the astronomical *error* function of the 19th century – of these objects are described by a two-dimensional Gaussian. The probability of two objects being counterparts to one another is therefore, as required, a function of both the separation between the two sources' positions, and a combination of the level of precision to which these observations can be known. Also necessary, the Gaussian distribution allows for the rejection of all potential counterparts, and allows for the acceptance of an object that is not necessarily the closest in sky separation. The AUF is the PDF that represents our belief as to the location of the object given its observed position.

Here the word *belief* is being used in its technical Bayesian sense, the epistemological interpretation of knowledge. In Bayesian inference, an initial *prior* belief is assigned to some hypothesis, and subsequent measurements of data update the belief in that hypothesis. The term *Bayesian* refers to the Reverend Thomas Bayes, and his essay solving a

problem set out previously by de Moivre (Bayes, 1763). So influential is this essay that the common result for the probability of a hypothesis given some measurements is referred to throughout the literature as "Bayes' Theorem" or "Bayes' Rule". However, the eponymous theorem was never stated in its canonical form anywhere in Bayes' work, and had been recognised as the product rule of conditional probabilities in previous works – indeed, de Moivre (1718) had already noted this in his *Doctrine of Chances* essays, on which Bayes (1763) was building! It took a further 11 years before Laplace (1774) generalised the result and applied it to the problem of inference. The term "Bayes' Theorem" is itself slightly troublesome as it suggests a pre-ordained, rigid methodology and an obvious solution to a certain kind of problem. It is therefore sometimes used in the literature without a deeper consideration of the problem being considered and the inferences being drawn. To quote the great statistician Edwin Thompson Jaynes, from his book *Probability Theory* (Jaynes and Bretthorst, 2003): "...the calculations we are doing – the direct application of probability theory as logic – are more general than mere application of Bayes' theorem; that is only one of several items in our toolbox".

A hundred years of investigation by at least a half dozen mathematicians and astronomers has led, from a seemingly unconnected series of investigations into the motions of the planets, to a crucial result in the history of the cross-matching process. We now have the mathematical language to describe the confidence we have in our beliefs, allowing for the formation of rigorous, quantifiable matching hypotheses. In fact, at this point one might consider the story to be at an end: we have the most advanced telescopes ever built, with the best detectors ever designed; we have the mathematics to describe the positions of our sources; and we have the framework to assign probabilities to our source identifications.

1.7 The Multiplicity of Counterpart Matching

However, there is still one additional philosophical contemplation that must be considered. Throughout history, the "catalogue matching" problem has often been complicated by

competition for assignment as the counterpart to one survey's source by multiple objects in a different catalogue. Whether this is the highly uncertain X-ray positions of the 1960s and 70s, or more recently with high angular resolution datasets such as the *Gaia* Data Release 1 (DR1; Gaia Collaboration et al., 2016b; Gaia Collaboration et al., 2016a), there are many instances where there is a many-to-one matching problem. These two sources of source multiplicity have very different implications. These effects have come under renewed scrutiny very recently, with extra care being taken with nearest neighbour matches to remove problematic identifications with clear multiplicity (e.g., Kellogg et al., 2015). Malkov and Karpov (2011) extend their cross-matching algorithm to include the ability to distinguish between single and binary stars, using the differences in the photometry of the two astrophysical hypotheses to break the otherwise astrometrically degenerate measurements. This problem is, more generally, an assignment problem, with work being undertaken to develop tools to determine the best pairing – or pairings – between one source in one catalogue and several objects in a higher angular resolution dataset (Budavári and Basu, 2016).

For the case of one set of uncertain data, there is typically a true one-to-one match that is simply unknown, such as in the X-ray matching problem. These sources are sparse enough on the sky to avoid origin confusion, but simply cannot be pinpointed to the accuracy required to know their position uniquely in the (e.g.) optical dataset, with much higher density and lower average spacing between sources. I have already discussed this type of multiple counterpart matching issue, focussing on the X-ray matches in particular in Section 1.2. While the problem has lessened over the years with technological improvements, with a period of relative matching ease highlighted in Section 1.4, it remains a fundamental limit for some types of survey. X-ray source cross-matching will still always require some level of source follow-up, with López et al. (2017) following up ultraluminous X-ray sources with the *William Herschel* Telescope observations, for instance.

The second cause of multiplicity is much harder to overcome. The physical limitations of the resolution power of telescopes decreases with longer electromagnetic radiation

wavelength, and thus sub-millimetre wavelength datasets intrinsically suffer from point-spread function (PSF) beam spreading (e.g., Sato et al., 2002). This causes the light from multiple sources to blend together, with detector counts often being difficult to assign to any given source. This limitation cannot be avoided, although it can be overcome with telescope size, and interferometry. This is why sub-mm and radio telescopes are some of the largest telescopes in the world, with the James Clerk Maxwell Telescope being 15 metres in diameter, and the Robert C. Byrd Green Bank Telescope, the world's largest steerable telescope, being 100 metres across. Interferometry, on the other hand, uses multiple telescopes in combination to increase the effective telescope diameter beyond that of any individual telescope in the configuration. Thus the Atacama Large Millimeter Array, consisting of 66 telescopes with individual diameters of up to 12 meters, can operate as an effective single telescope with a diameter of 16 kilometers. This unavoidable resolution limitation means that the *Herschel Space Observatory*, perhaps one of the most sensitive sub-mm telescopes to date, still requires meticulous care when assigning its sub-mm detections to counterparts in other wavelengths (Bourne et al., 2016).

In cases where the PSF of a telescope is large, either due to its longer wavelength, smaller diameter, or atmospheric effects causing poor seeing, the blending of sources can become troublesome. There are several studies on the effect the blending has on the sources detected throughout the literature of many different surveys, such as for any cosmic microwave background data obtained by potential future missions (Curto et al., 2013), and the contamination of gamma-ray detections from the *Fermi Gamma-Ray Space Telescope* (Daylan, Portillo, and Finkbeiner, 2017). Additionally, there is research conducted into the best way to assign catalogue counterparts when the two surveys' resolutions differ significantly. Perhaps the most obvious example, especially in recent years, is the release of the *Gaia* DR1 dataset. With a 0.1 arcsecond PSF, *Gaia* provides all-sky optical observations with an order-of-magnitude higher angular resolution than previous typical ground-based surveys. Marrese et al. (2017) provide a method for the matching of these high angular resolution *Gaia* data with several key archival catalogues,

tuning their methodology under the knowledge that there will be several *Gaia* sources to one counterpart, with *Gaia*'s ability to resolve much finer detail on its observations.

On smaller scales the inspection of high angular resolution data is still a good way to confirm whether there is significant blending in the observations of large PSF surveys. Perhaps the telescope for which blending is most often considered, *WISE* suffers from a large (approximately 8 arcsecond diameter) PSF due to its 40-cm diameter aperture. Operation at mid-IR wavelengths and all-sky coverage mean that it is incredibly powerful for the study of IR excesses, indicative of dusty discs around stellar objects. However, the blending of two sources, one with its own significant IR flux, could be misinterpreted by such a search. Therefore, such as in the case of the White Dwarf search conducted by the WIRED team (Debes et al., 2011; Dennihy et al., 2017), detections with significant *WISE* IR excess are visually inspected for additional detections in higher angular resolution surveys – in this case the 4-metre Visible and Infrared Survey Telescope for Astronomy (VISTA) in Chile. Similarly, the study using *WISE*'s *NEOWISE* reactivation (Mainzer et al., 2014) for long-baseline proper motion analysis confirmed candidate proper motion objects visually in the SDSS and 2MASS filter images to reject spurious double star blendings (Schneider et al., 2016). Morales and Robitaille (2017) analysed UKIRT Infrared Deep Sky Survey (UKIDSS) data to construct multiple source SEDs of *Spitzer* Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) observations. They found that, while approximately 90% of the GLIMPSE data had two UKIDSS sources in their SEDs, sources beyond the most dominant contributed relatively little to the overall flux in most cases.

Thus, on at least some level, the differing telescopes, conditions, locations, and wavelength ranges influence the intrinsic properties of a photometric catalogue. Whether caused by telescope size, or atmospheric effects, or long wavelength, different PSF sizes can cause the blending of multiple sources, a much more difficult multiplicity problem than the relatively simple case of a single detection with poor centroiding. This problem must be acknowledged and tackled, whether through quantitative methodology or time-

intensive follow-up, to avoid introducing unphysical effects into the conclusions drawn by any analysis of these datasets.

1.8 Summary

The cross-matching of different observations of the same area of the sky is not necessarily an easy job. The road to our current understanding of telescope optics, astrometry, and the formalism of match likelihood has been long, with a 200-year history. It is crucial that the images – whether they be photographic or electronic in origin – taken by both telescopes used to construct the catalogues being cross-matched be interpreted correctly, with their flaws, nuances and systematics known fully in order to truly know the forces of nature behind the distant source you wish to understand. I have laid out in this introduction a brief history of the attempts to better know these objects, and overcome the limitations and difficulties of astrophysical observations. We can now obtain detections of sources fainter than ever before, in higher detail than ever before. We can formally quantify the degree of confidence we assign to the pairing of differing wavelength detections, and we can overcome the nature-placed limits on the ability to see into the heavens by utilising multiple telescopes operating at multiple wavelengths, each with their own unique strengths and weaknesses.

In this thesis I explore the cross-matching of catalogues of astrophysical sources in more detail, expanding upon the problem of multiplicity and proposing new ways to improve the matching of catalogues with vastly different systematic effects, resolution, dynamic range, or wavelength coverage. In Chapter 2 I analyse the effect that the blending of multiple sources has on the astrometry of the dominant source. Chapter 3 proposes an improvement to the catalogue matching methodology, generalising the description of source association astrometrically and including a model-free description of source association photometrically. Chapter 4 combines the work of the previous two chapters, applying the work to create a reliable source of *Gaia* DR1 and *WISE* cross-matches. I discuss the method to include the blending of sources in the AUF of a photometric

catalogue source more rigorously and discuss the implications it has on other derived parameters, such as proper motions. I provide concluding remarks in Chapter 5, weighing the evidence as to whether the error box vs depth race is truly ever won. Table 1.1 provides a description of the symbols used in this thesis, while Table 1.2 gives the photometric quality flags used in each of the chapters. These flags allow for the selection of only high quality sources, as well as the removal of non-stellar sources where possible. The majority of these flags are selected using the documentation provided with the release of the given dataset; however, a few – namely those removing APASS catalogue entries and the minimum number of *Gaia* matches, both "matched" and "good" – are the result of either trial and error or a logical deduction criterion. I also thank Nigel Hambly for some excellent early advice on interpreting *Gaia* quality flags. Finally, Table 1.3 provides some background information on the different catalogues used in this thesis, such as wavelength coverage, catalogue size, all-sky coverage, and PSF size.

Table 1.1: Table showing the definition of symbols used throughout.

Symbol	Definition	Chapter
a, b	Semi-major and semi-minor star sky axes	2
A_ϕ	Counterpart PDF star area of consideration	2
A, $\mathcal{A}$	Area	1
b_ϕ	PDF of bright stars in A_ϕ	2
B	Magnitude density of sources at given magnitude	3
$c(m_\gamma, m_\phi)$	Symmetric counterpart magnitude PDF	2, 3
$c(m_\gamma \vert m_\phi)$	PDF of counterparts with magnitude m_ϕ	2
$C(m_\gamma \vert m_\phi)$	Integral of $c(m_\gamma \vert m_\phi)$ from $-\infty$ to m_γ	2
D	Differential source counts	3
$\mathrm{d}x$, $\mathrm{d}y$	Small sky widths defining sky cell area	2
$\mathrm{d}m$	Small range of stellar magnitudes	2
$f_\phi(m_\phi)$	Unmatched catalogue ϕ star PDF	2, 3

Continued on next page

Table 1.1 – continued from previous page

Symbol	Definition	Chapter
$F_\phi(m_\phi)$	Integral of f_ϕ from $-\infty$ to m_ϕ	2
F	Flux ratio of bright and faint objects	1
F_{contam}	Average flux contamination ratio of sources	3
$g(x_\gamma, x_\phi, y_\gamma, y_\phi)$	PDF of two stars being counterparts given offset	2
$G(\Delta x, \Delta y)$	PDF of two counterparts being offset in x and y	2, 3
h, h_ϕ	Astrometric uncertainty function of catalogue ϕ	1, 2, 3
i, j, k, l	Indices	2
K	A normalisation	2
l, b	Galactic sky coordinates	1, 2, 3
M	Total number of counterparts	1
m	Magnitude of bright object	1, 2, 3
m_i	Magnitude of differential source count break	3
N, N_i	Geometric number density normalisation constants	1, 3
N_{c}	Counterpart number density	2, 3
N_ϕ	Unmatched catalogue ϕ number density	2, 3
n_ϕ	Number of detected objects in catalogue ϕ	2
O	A normalisation	2
p_ϕ	PDF of all stars in catalogue ϕ	2
P_{match}	Counterpart match probability	3
P_{contam}	Probability of source being contaminated	3
Q	Contamination figure of merit	1
$\mathcal{R}_Y$	Radius defining circular integral	2, 3
r	Radial distance	1, 3
R	PSF radius	1, 3
s, t	Indices	2

Continued on next page

Table 1.1 – continued from previous page

Symbol	Definition	Chapter
T	Number of stars in a given magnitude range	2
U, V	Number of objects in circle of given radius	1, 3
W	Average number of PSF contaminants	3
x, y	Cartesian coordinates	1, 2, 3
Y	Fraction of integral	2, 3
z, z_i	Scaling for increase in star counts with magnitude	1, 3
$Z_{c\phi}$	Fraction of stars with counterparts	2
Z_ϕ	Fraction of stars with at least one star inside A_ϕ	2
α, δ	Celestial Coordinates	1, 2, 3
γ	A catalogue	2, 3
Δm	Given magnitude offset from central source	1, 3
$\Delta m_{\max}$	Maximum magnitude offset	3
Δr	Width of radial annulus	1
ϵ	A catalogue	2
ζ, λ	Sets of catalogue detections	2
η	Photometric likelihood ratio	2, 3
θ	Position angle of sky axes	2, 3
μ_α, μ_δ	Proper motion in sky coordinates	1
ξ	Astrometric likelihood ratio	2, 3
ρ	Correlation of celestial sky axis uncertainties	2, 3
σ, σ_α, σ_δ	Celestial sky axis uncertainties	1, 2, 3
σ_{quoted}	Astrometric uncertainty given in catalogue	1
σ_{core}	Uncertainty fit to the inner radius of an AUF	1
ϕ	A catalogue	2, 3
ψ	A contamination hypothesis	3

Continued on next page

Table 1.1 – continued from previous page

Symbol	Definition	Chapter
ω	A contamination hypothesis	3
$\frac{\mathrm{d}N}{\mathrm{d}r}$	Number of separations per unit distance	1
$\frac{\mathrm{d}N}{\mathrm{d}A}$	Number of stars per unit area	1
$\frac{\mathrm{d}N}{\mathrm{d}A\,\mathrm{cat}}$	Number of detected sources per unit area	1

Table 1.2: Table showing the various flags for non-stellarity, detection and photometric quality for the catalogues used. In cases where flags refer to a specific filter, *Gaia* only uses the G filter; IPHAS uses the r and i filters; 2MASS is cleaned using the J, H, and K_s filters; *WISE* is $W1$, $W2$, $W3$ and $W4$, APASS uses B, V, g, r, and i; and *Spitzer* is cleaned using the [3.6] and [4.5] passbands.

Catalogue	Flag	Criteria	Chapter
Gaia	Non-stellar	astrometric_excess_noise > 0.865mas and astrometric_excess_noise_sig > 2	1, 2
	Low Quality	astrometric_excess_noise > 0.865mas and astrometric_excess_noise_sig ≤ 2; or astrometric_n_good_obs_al + astrometric_n_good_obs_al < 60; or matched_observations ≤ 8	
	Non-Stellar	astrometric_excess_noise > 2.375mas and astrometric_excess_noise_sig > 2	3
	Low Quality	astrometric_excess_noise > 2.375mas and astrometric_excess_noise_sig ≤ 2; or astrometric_n_good_obs_al + astrometric_n_good_obs_ac < 60; or matched_observations ≤ 8	
WISE	Non-stellar	"Contam" flag is either "D", "P", "H", or "O"; or "ext" flag is 2, 3, 4, or 5	1, 2, 3
	Outside Dynamic Range	"Phqual" flag is "X" or "Z"; or "detbit" == 0; or Mag == NaN; or "sat" flag > 0; or σ_{Mag} == NaN	
	Low Quality	"Phqual" flag is "U"; or "Contam" flag is "d", "p", "h", or "o"; or "ext" flag is 1; or "var" flag is > 5 or "n"; "nblend" flag is > 3; or "moonlev" >"1"	

Continued on next page

Table 1.2 – continued from previous page

Catalogue	Flag	Criteria	Chapter
Spitzer	Outside Dynamic Range	Saturation Flag is set, or Artefact of Wing Saturation flag is set	3
	Low Quality	Dark Current flag is set, Flat Field flag is set, Latent Image flag is set, Bad Pixel flag is set, In-band or Cross-band Merge Confusion flags are set, or Edge of Frame flag is set	
APASS	Outside Dynamic Range	Mag > 20 or Mag < 10	1
2MASS	Non-stellar	"Galcontam" or "Mpflag" flags set	1, 2
	Outside Dynamic Range	"Blend" flag == 0; or "Read" flag == 0 or 3; or Mag == NaN; or σ_{Mag} == NaN	
	Low Quality	"Photqual" flag is not "A", "B", or "C"; or "Read" flag is not 1 or 2; or "Blend" flag is not 1, 2, 3; or "Contam" flag is not "0" or "c"	
IPHAS	Non-stellar	$p_{\mathrm{star}} < 0.9$	1, 2
	Outside Dynamic Range	Mag == NaN; "Saturated" flag set; or σ_{Mag} == NaN	
	Low Quality	"Deblend" or "BrightNeighbour" flagged; $\sigma_{\mathrm{Mag}} > 0.1$; or $\lvert\mathrm{Mag} - \mathrm{AperMag1}\rvert > 3\sqrt{\sigma^2_{\mathrm{Mag}} + \sigma^2_{\mathrm{AperMag1}}} + 0.03$	

Table 1.3: Table showing some background information on the various catalogues used throughout this thesis. [1]The *Spitzer* survey here is the GLIMPSE survey, itself a composite of the GLIMPSE I, II, 3D, and 360 surveys, as well as various smaller observations of the Galactic plane. [2]Aperture photometry used, PSF width unknown.

Catalogue	Filter	Wavelength Region	Effective Filter Wavelength	Catalogue Size	All Sky?	PSF FWHM
Gaia	G	Optical	600nm	1.14 billion	Y	0.1 arcsecond
WISE	$W1-4$	Mid-IR	3.37 - 22.19μm	747 million	Y	6.1 arcsecond
Spitzer[1]	[3.6], [4.5]	Mid-IR	3.6 - 4.5 μm	151 million	N	2.0 arcsecond
APASS	$BVgri$	Optical	440 - 760 nm	61 million	Y	N/A[2]
2MASS	JHK_s	Near-IR	1.24 - 2.16 μm	455 million	Y	2.9 arcsecond
IPHAS	ri	Optical	620 - 760 nm	219 million	N	1.1 arcsecond

Chapter 2

The Effect of Unresolved Contaminant Stars on the Cross-Matching of Photometric Catalogues

Were the succession of stars endless... there could be absolutely no point, in all that background, at which would not exist a star.
— Edgar Allan Poe, Eureka (1848)

2.1 Introduction

Broadband photometry is a staple of astrophysics, able to provide a wealth of information on a plethora of objects of interest without the time requirements of spectroscopy. To break degeneracies in theoretical models and gain as much understanding as possible, oftentimes multi-wavelength coverage is required. This means combining the efforts of several surveys, where teams and collaborations have independently taken photometric images of the sky in various wavelength regimes. It is therefore of vital import that the same stars in separate catalogues are correctly identified. Traditionally, the method for matching two catalogues together uses the smallest distance between a given star in one catalogue and stars in the opposing catalogue, pairing those stars that both have the other

star as their closest corresponding star. Additionally there is a cutoff radius beyond which no pairs can be matched, typically 2 or 3 arcseconds.

Recently, the idea of matching between catalogues following a probabilistic approach (starting with Sutherland and Saunders, 1992; see Section 1.5 for a more comprehensive discussion) has become common. It gives a more flexible approach by adjusting the size scale over which matches are considered likely to match the precision of the detections. High quality, precise astrometric data only allow matches between stars close to one another, while less precise data are allowed to have counterparts beyond the 2-3 arcsecond typical maximum nearest neighbour cutoff.

Nearest neighbour matching is equivalent to carrying out probability-based matching using a "top-hat" function with the cutoff radius, inside which a star is equally likely to exist at any distance from another detection and outside which it is impossible to be matched. Astrometrically the full probability-based method is favourable because the top-hat is unphysical. To improve upon this "top-hat", a more complete description of the probability of detecting the counterpart in the opposing catalogue at a given separation is required. These probabilities of star pairs being counterparts to one another as a function of separation are themselves a function of what I shall refer to as the astrometric uncertainty functions (AUFs; Section 1.6). Usually, these distributions are assumed to be purely Gaussian. This does not account for any wings to the distributions themselves, yet these are known to exist (see, e.g., Krawczyk et al., 2013 Figure 4 or Munari et al., 2014 Figure 2). The assumption that the AUF is Gaussian could lead to a significant mis-identification of a large number of counterparts. In the probability-based matching case this incorrect matching is due to the assumed shape of the distributions not being a good description. In the nearest neighbour matching case it is caused by the accepted cutoff radius being too small.

Probability-based matching also has increased flexibility in allowing for comparisons between two detections in one catalogue by including additional information, such as magnitudes (e.g., Budavári and Szalay, 2008 and Naylor, Broos, and Feigelson, 2013).

If two stars are close enough to the same star in another catalogue to be considered likely matches, the extra parameter space allows for the possibility of rejecting an unfavourable match that is serendipitously nearer than the better match. However, this extra information can not be used if the AUFs are ill-defined, so it is vital that they are correct. As I discussed in Section 1.7, oftentimes in catalogue cross-matching there is one catalogue of higher resolution than the other, leading to a multiplicity of potential counterpart matches. It has been known for a while that the effects of telescope point spread functions (PSFs) cause faint sources to be underrepresented in counts, blocked out by the light from the more dominant sources in the sky (e.g., Naylor, Broos, and Feigelson, 2013). However, little literature exists on how these fainter sources might influence the central sources blocking their detection.

In this chapter I will explain how crowding in high density regions causes long, non-Gaussian tails in the AUFs. I will begin by initially introducing the catalogues being used throughout the chapter in Section 2.2, and in Section 2.3 defining the AUF more formally. I will then examine the spatial distribution of an examples of matches for a crowded region of the Galactic plane before discussing some possible reasons for the non-Gaussianity seen in the distributions, concluding that they cannot satisfactorily explain the results in Section 2.4. I introduce the effect of crowding seen in photometric catalogues in Section 2.5. This is used to explain how this effect causes the non-Gaussian tails, before I test the hypothesis with some simple approximations in Sections 2.6 and 2.7. I then put the effect into context for several additional large scale, commonly used surveys in Section 2.8. Finally, I offer some options to overcome the issue of contamination in Section 2.9. Here I give some cases where one can maximise the number of true matches at the expense of false positives, or, alternatively, minimise the number of false positives and contaminated matches. I define symbols used throughout the chapter in Table 1.1.

2.2 Catalogues

The matching of photometric catalogues has significant problems in very crowded fields, and is at its worst in the Galactic plane, especially towards the Galactic centre. In addition, the crowding becomes more problematic with increasing seeing or larger PSFs. The crowding of stellar fields is then a function of both stellar density and PSF area, which is why I have chosen to focus on *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al., 2010) for most of this chapter. With a $\simeq$ 6 arcsecond full-width at half maximum (FWHM) in bands $W1 - W3$ and a relatively deep survey reaching $W1 \simeq 17$, the *WISE* datset suffers from significant crowding. At the other extreme, the recently released *Gaia* Data Release 1 (DR1; Gaia Collaboration et al., 2016b; Gaia Collaboration et al., 2016a) provides excellent and unprecedented astrometric precision, and with a $\simeq$ 0.1 arcsecond FWHM should be effectively uncrowded.

Initially I will consider *Gaia* and *WISE*, but I will introduce the AAVSO Photometric All Sky Survey (APASS; Henden and Munari, 2014), INT Photometric Hα Survey (IPHAS; Drew et al., 2005; Barentsen et al., 2014), and Two Micron All Sky Survey (2MASS; Skrutskie et al., 2006) in a later section. To ensure minimal erroneous or poor data in the catalogues, I first clean them to remove either known non-stellar sources, or to remove spurious, low-quality, saturated, and upper flux limit objects, as detailed in Table 1.2.

2.3 The Astrometric Uncertainty Function

The probability that two stars in two photometric catalogues are counterparts to one another is the probability that the stars from the two catalogues are drawn from the same original sky position, involving the AUFs of both catalogues. However, the order-of-magnitude higher precision in the *Gaia* dataset simplifies the problem such that the probability of matches reflects only the uncertainties in the second catalogue. Thus, only the AUF of *WISE* detections in this instance is required.

This means the probability of measuring a source, with "true" position at the origin, at position x, y can be modeled as a centered, circular, two-dimensional Gaussian (Quetelet, summarised by Herschel, 1857)

$$h(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \tag{2.1}$$

where σ is the astrometric uncertainty in either of the orthogonal axis directions. The astrometric uncertainty can be approximately related to the photometric signal-to-noise ratio (SNR) and image PSF scale length. King (1983) quotes the relationship as the scale length of the image divided by the SNR.

When considering a circular geometry, this form can be transformed to radial coordinates by integrating over θ, which changes the Gaussian distribution to a Rayleigh distribution, given by

$$h(r, \sigma) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{2.2}$$

$h(x, y, \sigma)$ is a probability density function, the probability per unit area, that the *WISE* star will be detected at an offset x, y from the *Gaia* source. Alternatively, $h(r, \sigma)$ is the probability per unit length that the *WISE* star is detected at a radial offset r from the *Gaia* source. It is the function $h(r, \sigma)$ that I will compare to the data in Section 2.4.

2.4 Fitting the Distribution

To check the validity of h, the AUF, it must be tested against some example data. Consider a large sample of matches, i.e. pairs of stars, all of which have a similar astrometric uncertainty σ. The number of matches per unit distance in a narrow annulus r to $r + \Delta r$ is

$$\frac{\mathrm{d}N}{\mathrm{d}r}(r, \sigma) = \frac{M}{\Delta r} \int_{r}^{r+\Delta r} \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) \mathrm{d}r, \tag{2.3}$$

where M is the total number of matches. Assuming all stars in the sample are true matches (see Section 2.4.2 for further discussion), the expected number of stars per unit distance can be compared with the number detected.

In this section I will consider matches between *WISE* and the *Tycho-Gaia* Astrometric Solution (TGAS; Michalik, Lindegren, and Hobbs, 2015) for an 800 square degree region of the Galactic plane ($100 \leq l \leq 140$, $-10 \leq b \leq 10$). Although the TGAS is a relatively bright subset of the full *Gaia* DR1 dataset – with the cleaned dataset containing ≈73000 stars in contrast to the ≈15.6 million *WISE* sources in its cleaned catalogue in the same region – limiting the match numbers, the proper motions will be required, which are only available for TGAS stars, in Section 2.4.2. I will discuss the effects of the full magnitude range in Section 2.8, and find the magnitude cut does not affect the conclusions drawn in this section.

2.4.1 Uncertainties for *WISE* Data

Matching between the two catalogues, *WISE* stars are taken in a narrow range of σ values (typically $\lesssim 0.01$ arcsecond) and nearest neighbour matched to the TGAS dataset. From this the number of stars in given radius bins is found, and the number of stars per unit radius within each annulus plotted, along with the assumed astrometric distribution, based on the quoted uncertainties. Figure 2.1 shows the resulting distribution for one narrow range of uncertainties $\sigma = 0.039 \pm 0.001$ arcsecond. It can be seen that the distribution is reasonably well described by a Rayleigh distribution in the inner region, below $r \simeq 0.1$ arcsecond, but that there is a significant non-Gaussian tail to the distribution of match distances.

2.4.2 Common Sources of Additional Astrometric Sources

There are two obvious potential causes of non-Gaussian data: a population of uncorrelated false matches, and the effects of proper motion on the apparent match distance between two catalogues of different epochs. As shown below, neither of them can adequately

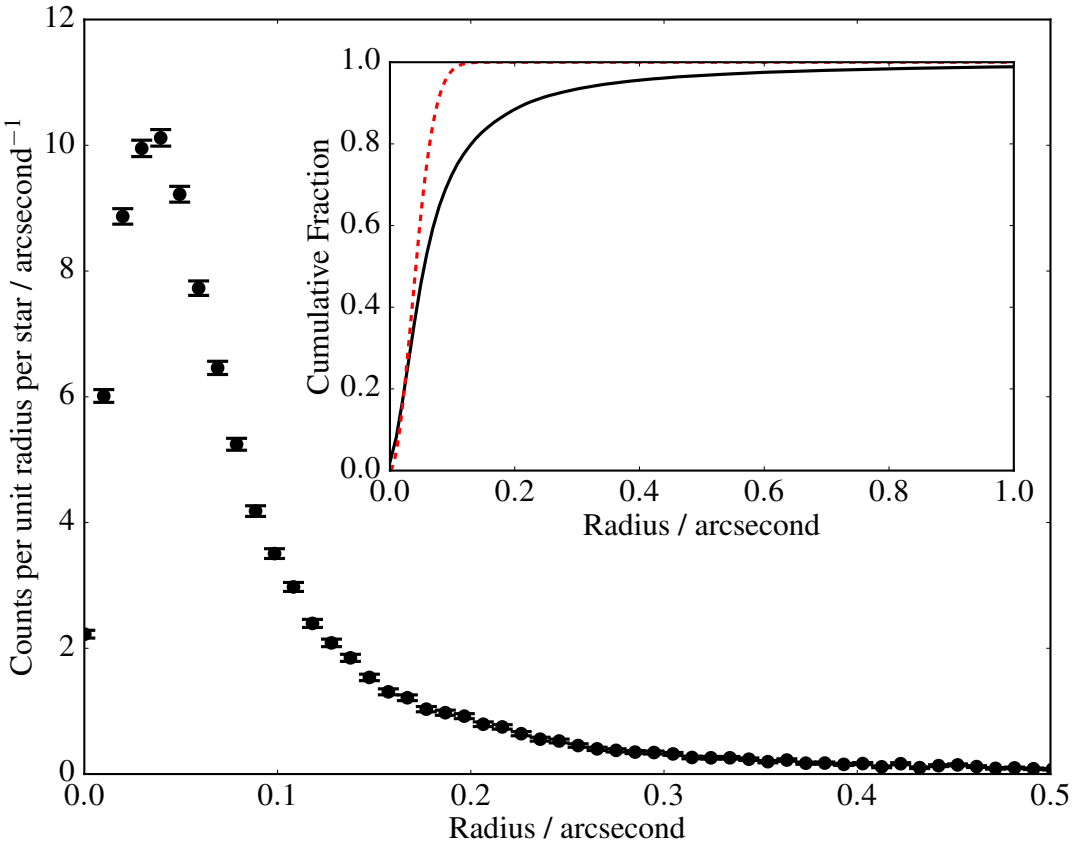


Figure 2.1: The separation of nearest neighbour matches between TGAS and *WISE*, for *WISE* objects with quoted uncertainty $\sigma = 0.039 \pm 0.001$ arcsecond. The inset Figure shows the cumulative distribution, with reference cumulative Rayleigh distribution of $\sigma = 0.039$ arcsecond shown as a red dashed line.

explain the effect entirely, requiring an alternative explanation.

2.4.2.1 Proper motions

Proper motions are often cited as being the cause of these "wings" at large separations (e.g., section 6.4 Figure 2 of Cutri et al., 2012; Appendix A1 of Flesch and Hardcastle, 2004). As *WISE* operated in 2010 while *Gaia* records positions in epoch J2015 it must be checked whether this is a significant cause of match offsets. The *Gaia* proper motions in the orthogonal axes for all stars in the 800 square degree region of the Galactic plane used to construct the distributions in Figure 2.1 were obtained.

The new celestial coordinates for the *Gaia* positions, transformed from the J2015 epoch to *WISE*'s J2010 epoch were calculated as

$$\alpha_{\rm new} = \alpha - 5{\rm year} \cdot \mu_\alpha \left[\cos(\delta)\right]^{-1}, \tag{2.4}$$

with an equivalent transformation for declination, where μ_α and μ_δ are the projected proper motions in the two orthogonal sky axis directions. The new distribution of proper motion-corrected separations was compared to a Gaussian of the average uncertainty $\sigma = 0.039$ arcsecond, shown in Figure 2.2. As can be seen, while the distribution tightens slightly towards smaller separations, the large, non-Gaussian tail remains beyond $r \simeq 0.1$ arcsecond. This leads to an incompatible cumulative distribution shown inset to Figure 2.2. The non-Gaussian tail increases with decreasing brightness (see Section 2.7 for more details), and the average magnitude of stars in Figure 2.2 is bright, at W1$\simeq$ 11. As such, most of the non-Gaussianity of the distributions cannot be explained with proper motions.

2.4.2.2 Uncorrelated False Matches

While the non-Gaussianity to the match distributions cannot be explained with proper motions, these are purely nearest neighbour matches. Some contamination from uncorrelated stars is expected which could potentially explain the non-Gaussian wings. At its most dense, there are 2×10^4 *Gaia* stars per square degree in the Galactic plane region in

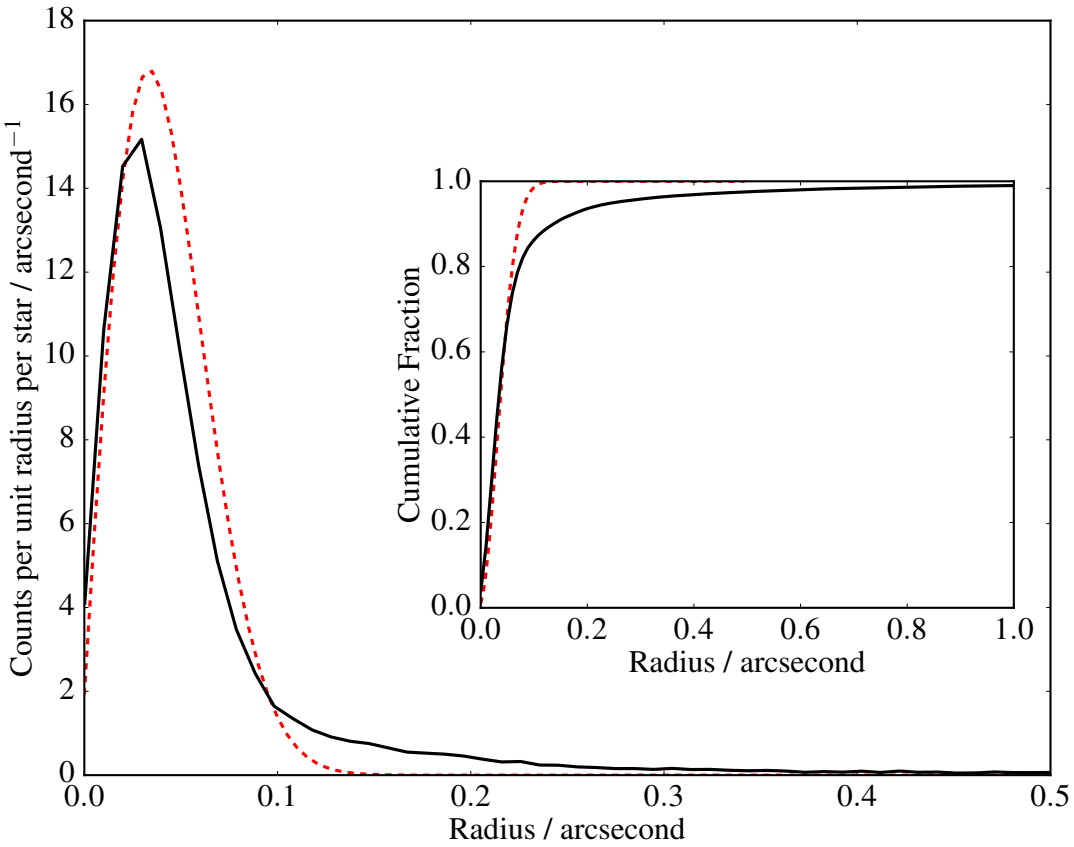


Figure 2.2: The effects of proper motions on *WISE*-TGAS matches with *WISE* astrometric uncertainty $\sigma = 0.039 \pm 0.001$ arcsecond. The distribution of separations, corrected for proper motion during the five year gap between observations, is shown as a solid black line. These are compared to the expected Gaussian of uncertainty $\sigma = 0.039$ arcsecond, shown in the red dashed line. The proper motion correction fails to account for most of the matches seen at large separations in the non-Gaussian tail.

question. The expected number of randomly placed objects in a circle of a given radius, U, is the multiple of the stellar density, $\frac{\mathrm{d}N}{\mathrm{d}A}$, and the area, A,

$$U = \frac{\mathrm{d}N}{\mathrm{d}A} \times A = 2 \times 10^4 \,\mathrm{deg}^{-2} \times \pi \left(\frac{0.5 \,\mathrm{arcsecond}}{3600 \,\mathrm{arcsecond/deg}} \right)^2 = 0.0012, \tag{2.5}$$

where, as per Figure 2.1, the circle has been limited to a radius of 0.5 arcsecond. Therefore 0.1% of the stars are expected to be false matches. These numbers are upper limits, as the nearest neighbour scheme employed reduces contamination beyond the radius of the true match separation for each star. It simply must be concluded that the distribution wings cannot be explained with uncorrelated star contamination.

2.5 Explaining the Distribution Wings

2.5.1 Star Spatial Distributions

To explain the distribution of matches between the two catalogues, it is illuminating to consider a *Gaia* source of magnitude $15 \leq G \leq 15.25$. The offsets from this star to all *WISE* objects with radial offset <30 arcseconds can be found. Repeating this calculation for all such stars in a 25 square degree region of the Galactic plane at $120 \leq l \leq 125$, $0 \leq b \leq 5$ a density of *WISE* sources astrometrically near *Gaia* sources in a narrow *Gaia* magnitude range as a function of radial distance is built up, shown in Figure 2.3.

There are three distinct regions. First, beyond 10 arcseconds from the *Gaia* objects there is a constant density of sources, which are uncorrelated, additional *WISE* objects. Second, there is a tight clustering of detections inside $r \lesssim 2$ arcsecond, which are the *WISE* detections corresponding to the *Gaia* objects. Third, there is a region 2 arcsecond $\lesssim r \lesssim 10$ arcsecond where randomly placed objects appear at a lower density than those at larger r.

However, non-match stars – those in the *WISE* catalogue whose G magnitude would

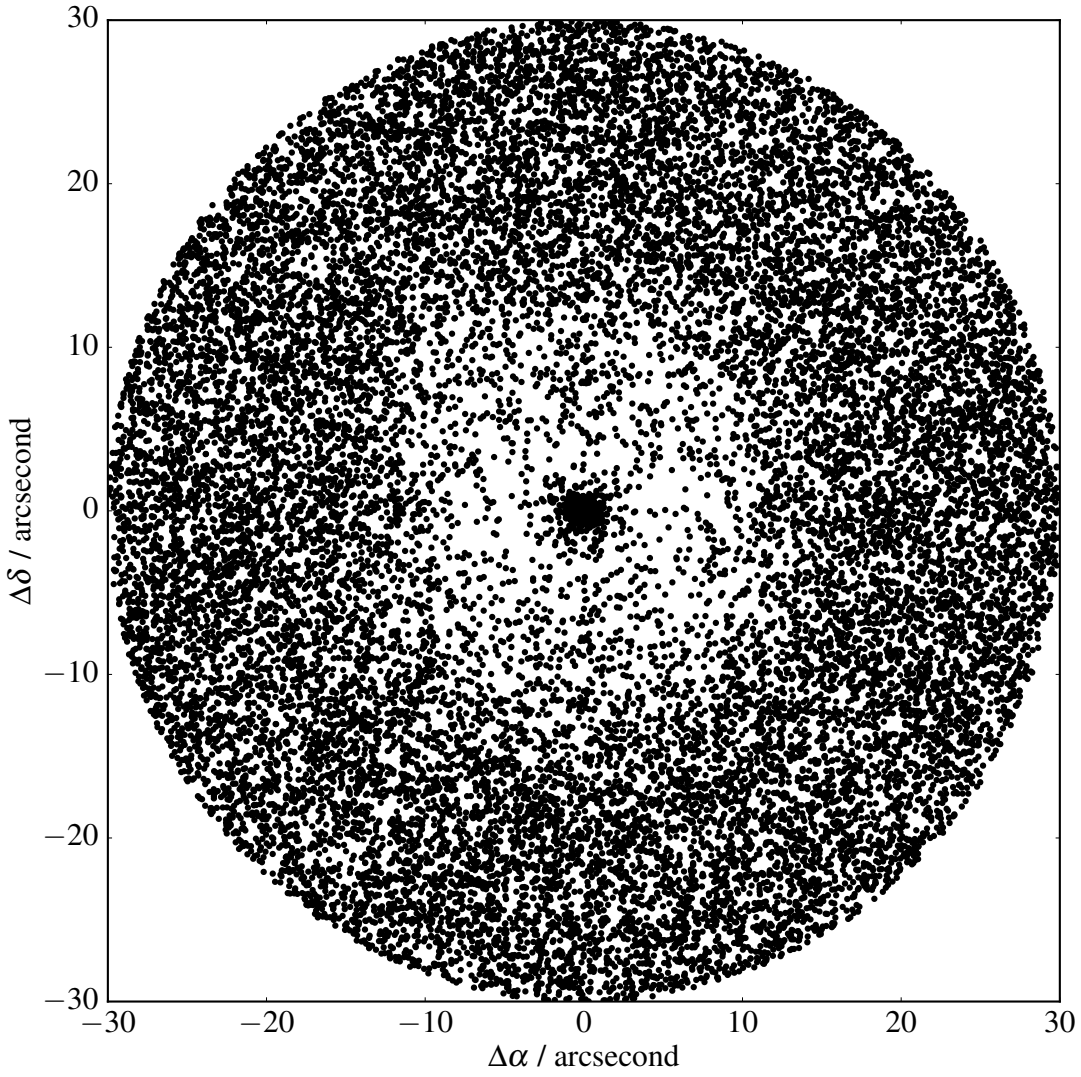


Figure 2.3: The spatial separation of all *WISE* stars within 30 arcseconds of *Gaia* sources $15 \leq G \leq 15.25$, for a $5^\circ \times 5^\circ$ slice of the Galactic plane. Background sources are seen at a constant density surrounding a clump of counterpart stars in the centre. However, the background density decreases within $\lesssim 10$ arcsecond due to the crowding out of the fainter background sources by bright counterparts.

lie outside of the 0.25 magnitude range – are not correlated with those stars that do lie in that small magnitude range. It is therefore expected that they have a constant stellar density across the entire sky, meaning that between 2 and 10 arcsecond radial distance the density of objects in some small area should be the same as beyond 10 arcseconds. This apparent reduction in stellar density is caused by crowding, a well known issue where bright sources dominate and cause non-detections of fainter objects inside their PSF, reducing the number of objects measured at these intermediate distances.

The important point to stress here is that these stars have not gone away – they are merely absorbed into the PSF of the bright star. This effect can be combatted by deblending multiple sources, in the form of either *active* or *passive* deblending. Passive deblending allows for the correction of blending between sources whose centroids are sufficiently separated to be resolved but whose PSF wings may introduce additional flux if not properly treated. In the case of *WISE*, sources must be at least 24 arcseconds apart to be passively deblended, but must also be within 2.5 magnitudes of the brightest source in the blend group. Active deblending occurs when the minimum reduced chi-squared for a single PSF model fit to a given source identified on a detector is above a critical threshold ($\chi^2_\nu > 1.5$), at which point the *WISE* pipeline attempts to fit a second component within the PSF. At most the pipeline is limited to one additional component from active deblending, putting an upper limit on the number of sources which can be deblended in crowded fields. The absorption of faint stars into the PSF of a brighter source causes flux contamination, which will compromise the photometry. However, since the vast majority of the contaminating sources will be objects significantly fainter than the main detection, with a low relative flux ratio, the photometric effect is small.

2.5.2 Contaminant Stars

More crucial, however, is the effect these sources have on the derived positions. Figure 2.4 shows an example schematic. A *Gaia* source and its true *WISE* match are offset by some small distance – on the order of tenths of arcseconds – but there lies inside the

$\simeq$ 10 arcsecond *WISE* PSF a second, undetected source with a tenth of the flux of the primary source, at $\simeq$ 3 arcsecond. This will tug on the position of the *WISE* primary by 0.3 arcsecond, changing the apparent separation between the *WISE* object(s) and the *Gaia* object. The distribution of separations – to be used for any potential probabilistic catalogue matching – is then a combination of two functions: the initial Gaussian-based statistics and the effects of undetected, embedded, contaminants.

There is a small but consistent thread in the literature highlighting the effect that source confusion – the inability to distinguish flux from one source from the flux of a second source – has on the properties of those sources. Olsen, Blum, and Rigaut (2003) discuss the effects of crowding on the next generation of Extremely Large Telescopes, while Jeong et al. (2006) account for the additional flux contamination in deep far-IR observations, and discuss strategies for dealing with confusion-limited detections. Hogg (2001) discusses this confusion in the context of artificial images, warning against using observations containing more than $1/30$ sources per beam. They postulate that these centroiding shifts may be the cause of the lack of optical counterparts to sub-millimetre data, much as is seen in Figures 2.2 and 2.4.

2.6 Validation with Synthetic Distributions

To test the effect these embedded stars could have on the AUF, I created a synthetic dataset based on simple geometric arguments. First the distribution of shifts that result when stars are contaminated within their PSF is required.

To obtain the shift distribution, test stars were placed inside 10^5 circles of a given sample bright star's PSF at random. These drawings assumed that the number density of stars increases by a factor of $z = 2$ with every step in magnitude. The flux-weighted position of the stars in each PSF was then found. Once all test contaminants had been drawn, the number of new positions in each given distance bin was recorded. Finally, the distribution was reduced to a probability density function by normalising the integral over all radii.

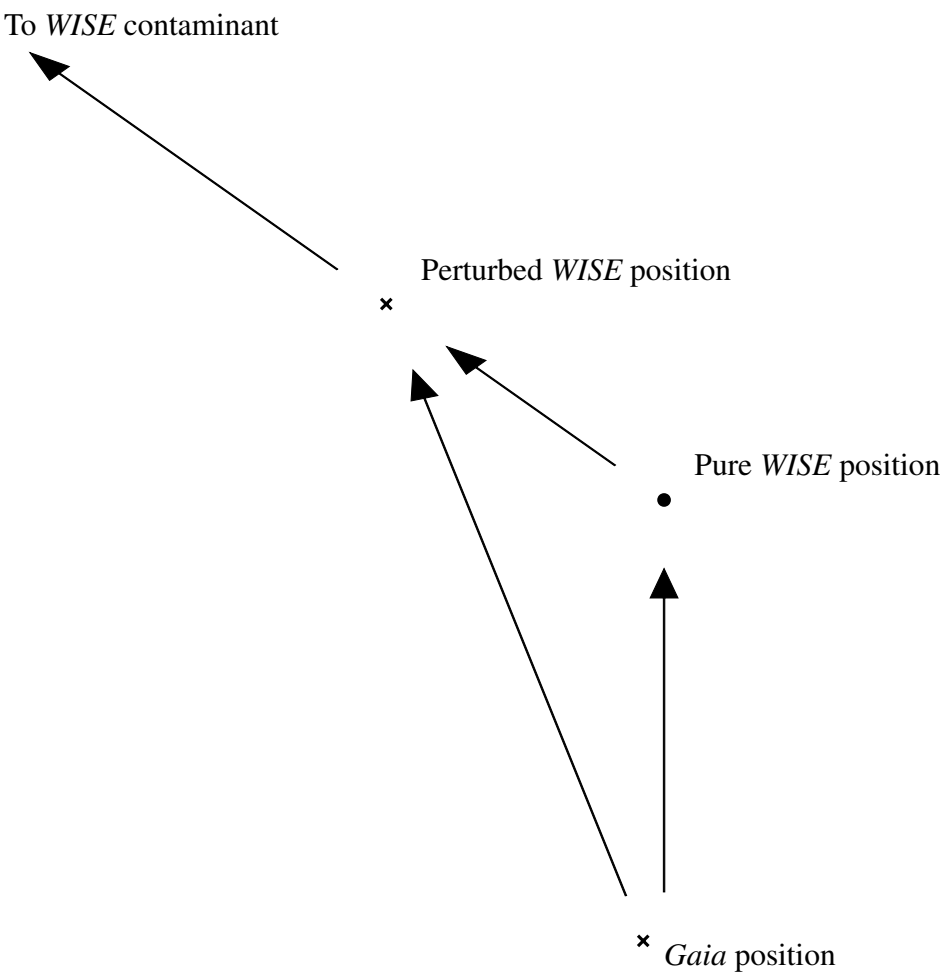


Figure 2.4: The effect of unresolved contamination on the measured position. Here, a *Gaia* object is separated from its true *WISE* counterpart by some distance. An undetected second *WISE* star within the *WISE* PSF causes the measured position to be shifted, causing a different separation to be calculated. This leads to a distribution of separations that is not merely based on Gaussian statistics.

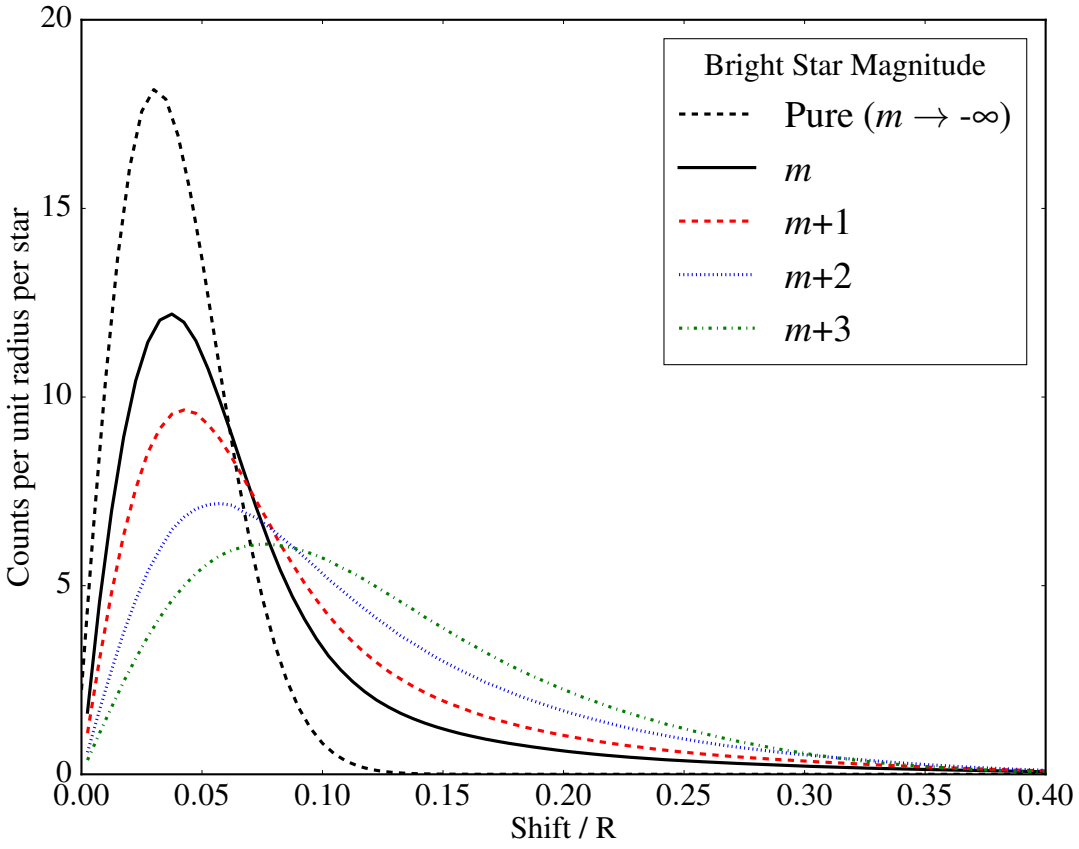


Figure 2.5: The effect of unresolved contaminating stars on distributions of synthetic positions in units of the PSF cutoff radius R. A Rayleigh distribution with σ = 0.05×FWHM was convolved with a derived contamination shifts distribution. The result is an inflation of the Rayleigh distribution uncertainty, as well as the introduction of the large, non-Gaussian tails similar to those seen in Figure 2.1, increasing with increasing stellar number density. The magnitudes, m through m + 3, represent increasing magnitudes of the central bright star, with a corresponding increase to the number densities of contaminants. The pure Rayleigh distribution, which effectively represents the contamination effects on an infinitely bright central star, is also plotted for reference.

The resultant function was convolved with a Rayleigh distribution of $\sigma = 0.05\times$FWHM, perhaps representing a star with SNR=20. The results of this are shown in Figure 2.5, for several bright stars with increasing magnitudes, representing increasing number densities of sky objects. The convolved functions still resemble the "pure" AUF in the inner region of the PSF, albeit with a broadened equivalent astrometric uncertainty, but the contamination also introduces a very long tail of separations. These objects are flux contaminated enough to introduce offsets on the order of 0.3-0.4×FWHM. This effect increases as the number density of objects increases, representing increased large separation contamination.

In summary, I suggest that the effect of astrometrically perturbed sources leading to large wings in distributions of counterpart distances, seen in the number of astrometric separations as a function of distance, is caused by the crowding out of fainter objects in the PSF. This leads to that fraction of stars – a very large fraction in regions of high stellar density, faint magnitudes, or large PSFs – with contaminant stars buried in their PSF exhibiting significantly non-Gaussian distributions in their detected positions. This will cause additional missed nearest neighbour matches if using a cutoff radius on the order of 1-2 arcseconds. It will also cause the resultant likelihoods derived from any probabilistic catalogue matching methods to fail in sampling the correct probability of matches and non-matches, also leading to a large fraction of false negative assignments.

2.6.1 Confirming the Numerical Contamination Shifts

To determine the contaminated distribution of astrometric positions, the distribution of shifts must first be calculated. It is this function that is subsequently convolved with a Gaussian, representing the pure position determination statistics. Given a bright star at the origin, a faint contaminant at radial distance r from the brighter object, with relative flux ratio F, will sit at a flux-weighted seperation s of

$$s = \frac{rF + 0 \times 1}{F + 1} = \frac{rF}{F + 1} = \frac{r}{1 + F^{-1}} = \frac{r}{q}. \tag{2.6}$$

Because $F \leq 1$, q will run from 2, when $F = 1$, upwards towards ∞ as $F \to 0$. Consider the number of contaminated bright objects, shifted to s, per unit area, $\frac{\mathrm{d}N}{\mathrm{d}A}$. This density can be found by integrating over all possible densities of stars of a given F,

$$\frac{\mathrm{d}N}{\mathrm{d}A} = \int_2^{q_0} \frac{\mathrm{d}N}{\mathrm{d}q\mathrm{d}A}\,\mathrm{d}q = \int_2^{q_0} \frac{\mathrm{d}N}{\mathrm{d}q\mathrm{d}\mathcal{A}}\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}A}\,\mathrm{d}q, \tag{2.7}$$

where $\mathcal{A}$ is the area subtended by the stars whose shifts will fall into smaller area A, the area of interest, with the upper integral limit q_0 being defined later. The first term in the integrand can be further expanded, as

$$\frac{\mathrm{d}N}{\mathrm{d}q\mathrm{d}\mathcal{A}} = \frac{\mathrm{d}N}{\mathrm{d}m'\mathrm{d}\mathcal{A}}\frac{\mathrm{d}m'}{\mathrm{d}q}. \tag{2.8}$$

The second term on the right-hand side of equation 2.8 requires the relationship between m' and q. Here $m' = m + \Delta m$, and Δm is defined as the magnitude offset of the fainter star relative to the bright central source, with $\Delta m \geq 0$. Δm is expanded as

$$\Delta m = -2.5\log_{10}(F) = 2.5\log_{10}(q - 1), \tag{2.9}$$

and thus

$$\frac{\mathrm{d}m'}{\mathrm{d}q} = \frac{2.5}{\log(10)(q - 1)}. \tag{2.10}$$

The number of stars per unit area per unit magnitude, the first term on the right-hand side of equation 2.8, is given by

$$\frac{\mathrm{d}N}{\mathrm{d}m'\mathrm{d}\mathcal{A}} = Nz^{m'} = Nz^m z^{\Delta m} = Nz^m z^{2.5\log_{10}(q-1)} = Nz^m(q-1)^{2.5\log_{10}(z)}, \tag{2.11}$$

where m is the magnitude of the central source.

Therefore the number of contaminating stars per unit q per unit area is

$$\frac{\mathrm{d}N}{\mathrm{d}q\mathrm{d}\mathcal{A}} = Nz^m \frac{2.5}{\log(10)}(q-1)^{2.5\log_{10}(z)-1} = N_0(q-1)^k, \tag{2.12}$$

with N_0 and k as simplifying constants as appropriate.

The next step is to consider the second term in the integrand on the right-hand side of equation 2.7. If, as assumed, the distance of a contaminating source relative to its apparent shift is q, then the relationship between area element $\mathcal{A}$, subtended by sources shifting the central object, and A, the area those shifts occupy, is

$$\mathcal{A} = q^2 A, \tag{2.13}$$

and therefore

$$\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}A} = q^2. \tag{2.14}$$

Finally, the integral has to be evaluated to q_0. As there can only be contamination inside the area of sky the bright object occupies, given by some PSF cutoff radius R, the lower flux ratio limit, or upper q limit, must be set such that $s = R/q$. As contaminants must increasingly move further away to produce the same apparent shift with decreasing flux ratio, this is the flux ratio where the object must lie on the edge of the defined crowding PSF circle. If the contaminant were any fainter, its required radial offset to produce the appropriate shift would place it outside the cutoff radius. Therefore $q_0 = R/s$.

Combining everything,

$$\begin{aligned}\frac{\mathrm{d}N}{\mathrm{d}A}(s) =& N_0 \int_2^{\frac{R}{s}} (q-1)^k q^2 \,\mathrm{d}q = N_0 \Bigg[\left(\frac{R}{s}-1\right)^{k+1} \Bigg(\left(\frac{R}{s}\right)^2 (k+1)(k+2) + \\ & 2\frac{R}{s}(k+1) + 2 \Bigg) - 4(k+1)(k+3) - 2 \Bigg] \times [(k+1)(k+2)(k+3)]^{-1}.\end{aligned} \tag{2.15}$$

This can be used to check the numerical simulations, plotting the probability density as

$$\frac{\mathrm{d}N}{\mathrm{d}s} = \frac{\mathrm{d}N}{\mathrm{d}A} \times 2\pi s. \tag{2.16}$$

The results are shown in Figure 2.6, albeit with the caveat that as the simulations only take into account flux ratios down to 1%, or $\Delta m = 5$, q_0 must be cut at the minimum of R/s and $1 + 1/10^{(-5/2.5)} = 101$.

When the central star is sufficiently bright that there are fewer than one relatively bright contaminant in each PSF the multiple star distribution agrees well with the single-star analytical solution, confirming the numerical simulations at bright central magnitudes. However, when the stellar density increases such that the contribution from more than one contaminant star of a sufficient flux ratio becomes large, the distributions do not agree anymore. There is therefore a need to use the numerical simulations when considering the effects of contaminant stars in the AUFs of photometric catalogues.

2.7 Quantifying the Contamination Levels

I showed that simple arguments about the effects of faint embedded stars inside brighter PSFs can reproduce similar results to those seen in the data (Figure 2.5 cf. Figure 2.1) in Section 2.6. However, the contamination levels from those faint stars must now be quantified. At a given stellar magnitude there will be some fraction of stars containing unresolved stars and another fraction which do not have within them additional sources. These are contaminated and uncontaminated objects, respectively. The uncontaminated fraction will still obey traditional Gaussian-based probabilistic statistics, but the contaminated stars will exhibit large shifts to their apparent position. This leads to the significant wings in their AUFs, as seen in, e.g, Figure 2.1.

The average number of stars inside the circle of the PSF can be calculated in a similar way to Equation 2.5, but using the radius of the circle the PSF subtends on the sky – typically 1-1.5 times the FWHM – for the area. In addition the number density is

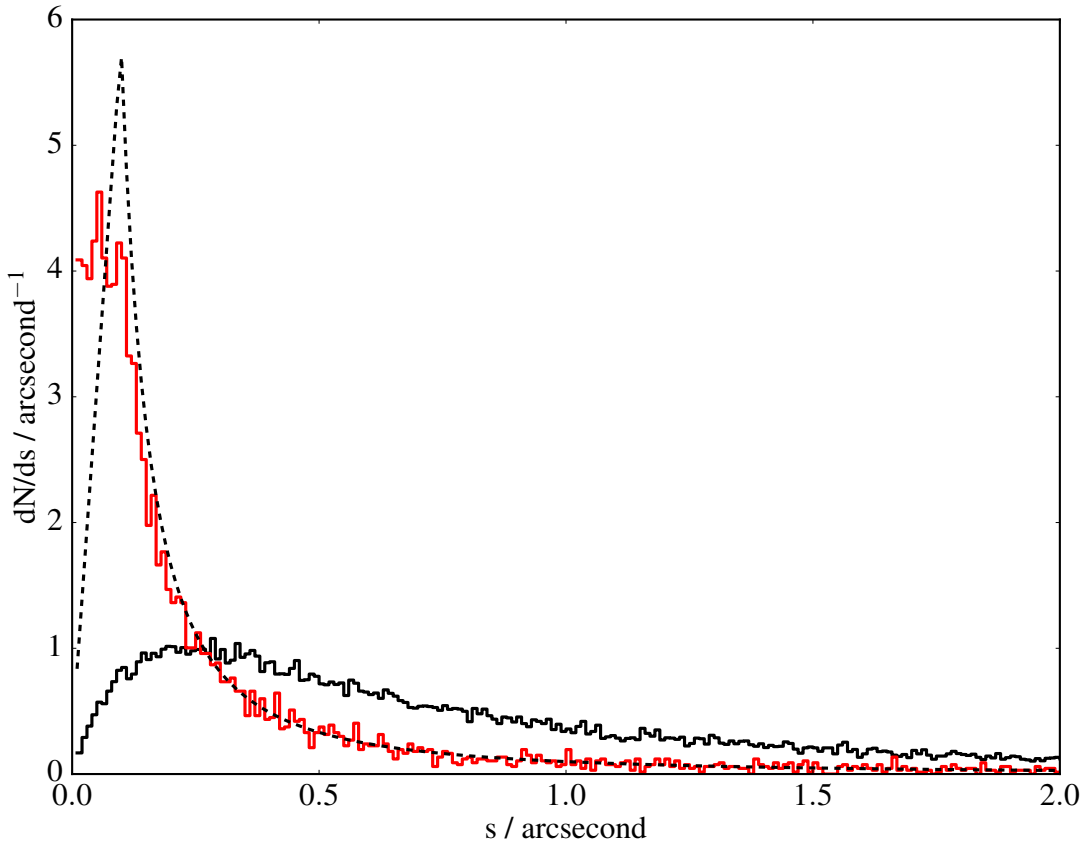


Figure 2.6: The distributions of contaminating shifts, for a theoretical star of $W1 = 14.95$ in black and $W1 = 9.5$ in red, where the histograms shows the numerical simulated data. The black dotted line shows the analytical single-star solution from Equation 2.15.

now the number of stars per square degree up to Δm magnitudes fainter than the star of magnitude m. This then gives a fraction of stars which are contaminated,

$$V = \frac{\mathrm{d}N}{\mathrm{d}A} \times A = \int_{m}^{m+\Delta m} N z^{m'} \, \mathrm{d}m' \times \pi R^2, \tag{2.17}$$

with R the PSF cutoff radius, N a normalisation factor, and $z \simeq 2$ the increase in stellar density with each step in magnitude. The choice of Δm is a reasonably arbitrary one, with stars technically being contaminated by faint stars with vanishingly small flux ratios, requiring an upper limit to the integral approaching infinity. However, the test data used in Section 2.6 show a convergence of the resulting AUFs for $\Delta m \gtrsim 4$. This suggests that above $\Delta m \simeq 4$, the distribution of contaminant shifts is dominated by the brighter contaminating stars, with very faint contaminants unable to affect the flux-weighted position. Thus $\Delta m = 5$ is a sensible choice, giving a flux ratio $F = 0.01$. For *WISE* in the Galactic plane, $l \simeq 120,\ b \simeq 0$, this gives a stellar density of $\simeq 6 \times 10^4$ deg^{-2} for $m = 13$; a factor of 3 increase over Equation 2.5. The contamination levels themselves use for the area in question $R = 10$ arcsecond, compared to the 0.5 arcsecond used when calculating the false positive rate.

Inside one out of every four PSFs of stars of $W1 \simeq 13$ there will be a star of $13 \leq W1 \leq 15$. This increases to approximately one star of $15 \leq W1 \leq 17$ inside the PSF of every 13th magnitude *WISE* star. Naturally some of these objects will be deblended during the reduction process, meaning that these numbers are upper limits, but as Figure 2.3 demonstrates, not all of them are successfully recovered, meaning they must be buried within the brighter detections.

2.7.1 The Contamination Figure of Merit, Q

The levels of contamination are dependent on the distribution of sources with magnitude and the size of the catalogue's PSF. To compare the contamination levels between catalogues requires a consistent metric.

Formally quantifying the stellar density requires fitting the number of stars per unit magnitude as a function of magnitude for the sky area in question. However, for a large fraction of the objects in the catalogue the contaminants that are perturbing their astrometry would be below the completeness limit of the catalogue, even outside of the bright star's PSF. This leads to the necessity of extrapolating the number density of sources below the completeness limit. It is more straightforward to just consider the stellar density of the overall catalogue, and assume the extrapolation of the number density to faint magnitudes.

Both a magnitude for which the contamination will be assessed (m) and a maximum acceptable contamination level must be decided on before the contamination levels between catalogues can be compared. For m, the median magnitude of the catalogue is a good choice, giving a lower bound to the contamination level of the fainter half of the catalogue. Additionally, the contamination level is fixed at 33%, the point at which a significant number of objects will be perturbed. These values then provide a baseline Q value, which can then be compared to values calculated for specific catalogues.

The number of stars per unit area in the magnitude range from the median magnitude of the catalogue to five magnitudes fainter is approximately ten times that of the detected source density. I showed in Section 2.7 that contaminants more than five magnitudes fainter than the central object do not contribute to the overall perturbation, and therefore limit the contaminants to $\Delta m = 5$. Choosing at most 33% of sources being contaminated, then

$$\int_{m}^{m+5} N z^{m'} \, \mathrm{d}m' \times \pi R^2 = \frac{\mathrm{d}N}{\mathrm{d}A} \times \pi R^2 = 0.33. \tag{2.18}$$

Substituting $R = 1.5\times\mathrm{FWHM}$ and $\frac{\mathrm{d}N}{\mathrm{d}A} = 10 \times \frac{\mathrm{d}N}{\mathrm{d}A\,\mathrm{cat}}$, where $\frac{\mathrm{d}N}{\mathrm{d}A\,\mathrm{cat}}$ is the source catalogue density gives

$$10 \times \frac{\mathrm{d}N}{\mathrm{d}A\,\mathrm{cat}} \times \pi \times (1.5 \times \mathrm{FWHM})^2 = 0.33. \tag{2.19}$$

This means that a 33% contamination level of stars of the median magnitude is

achieved when the contamination figure of merit

$$Q \equiv \frac{\mathrm{d}N}{\mathrm{d}A}_{\mathrm{cat}} \times \mathrm{FWHM}^2 \simeq 0.005. \tag{2.20}$$

It may be surprising that a catalogue where only a fraction of a percent of the sources might contain as contamination another source detected in the catalogue suffers from 33% perturbation. Howevever, the 0.5% result is simply the chance that a star above the completeness limit of the survey falls within a box with side length equal to the FWHM of the survey. The PSF length scale and, more importantly, the fact that stars are astrometrically perturbed by objects below the sensitivity of the survey both contribute to a much more significant contamination level. However, the Q value is a useful tool for comparing surveys of different spatial resolutions and dynmical ranges.

Additionally, the number of objects affected both photometrically and astrometrically throughout the dynamical range of the catalogue can be compared. Towards the bright end of the catalogue, the number density of stars contaminating is relatively low. Here any stars affected will have accurate astrometric positions, and so the undetected contaminants will lead to large astrometric offsets compared to their uncertainties. However, the fraction of stars affected is sufficiently small that the contribution to the AUF from contaminated stars may be negligible. At the faint end of the catalogue the opposite is true, where the effective stellar density is very high and therefore the fraction of stars photometrically compromised is high. However, the SNR rapidly decreases towards the completeness limit of the survey and thus the influence of the contaminant stars is diminished, lost amidst the inherent uncertainty in measuring the position. Astrometrically the most affected part of the catalogue is between these two extremes, in the region where the stellar density is still high enough to have a large fraction of stars contaminated, but with accurate enough positions that the effects of contaminants are easily detectable.

2.8 Surveys in Context and the Quoted-Core Distribution Uncertainty Relationship

While I have focussed mostly on the *WISE* AUF, it is salient at this point to mention how this effect changes the distributions of other catalogues. Here I will briefly discuss three additional, complementary, large-scale surveys: two optical surveys, APASS and IPHAS, and the near-IR survey 2MASS. These catalogues are especially useful as they allow for the direct probing of the effect of increasing stellar density and decreasing PSF scale length. I will also put *WISE* into a wider context.

I have shown evidence of a broadening of the AUFs relative to their assumed Gaussian positional uncertainties in Section 2.6. As a consequence, I fit the AUFs for large sections of the Galaxy for each survey in one square degree divisions, giving relationships between the quoted and best-fit Rayleigh distribution uncertainties. The relationship between the quoted uncertainty and best fit Rayleigh distribution is

$$\sigma_{\text{core}} = m\sigma_{\text{quoted}} + c, \tag{2.21}$$

with the core uncertainty such that the Rayleigh distribution best fits the smallest radial offsets of the given dataset, and the quoted uncertainty that as taken directly from their respective catalogues. I fit for some arbitrary offset c, but as expected the best fits have intercepts on the order $|c| \lesssim 0.05$ arcsecond, resulting in effectively a scaling between the quoted and core uncertainties.

However, as detailed further in Section 2.9, while these broadened Gaussian uncertainties are useful, it must be cautioned that these empirical uncertainties do not necessarily allow for the selection of uncontaminated objects. Figure 2.5 shows that there is significant overlap between the contaminated and uncontaminated distributions.

2.8.1 APASS

As an all-sky survey bridging the gap between the *Tycho-2* and SDSS surveys (Henden and Munari, 2014), APASS is a very important survey. However, it has a relatively large PSF, using a diameter of 15-20 arcseconds for its aperture photometry, and large detector pixels ($\simeq$ 3 arcsecond/pixel), leading to a significant fraction of contaminated stars and large wings in the APASS-*Gaia* separation distribution. This is mitigated slightly by its reasonably bright completeness limit, effectively reducing the stellar density at its faint end, giving a contamination fraction on the order of tens of percent, or a Q value of 3.4×10^{-3}.

APASS has very conservative astrometric uncertainties in DR9, requiring an empirical fit to any data being used in a probability-based matching process. In the Galactic plane (l $\simeq$ 120, b $\simeq$ 0) the core uncertainty is approximately 65% of the quoted uncertainty, decreasingly dramatically towards the Galactic pole ($b \geq 75$) where the core uncertainty is $\simeq$30% of the quoted uncertainty.

2.8.2 IPHAS

IPHAS used the *Isaac Newton* Telescope on La Palma to conduct a relatively large scale, deep survey of a section of the Galactic plane. The median PSF FWHM of $\simeq$ 1 arcsecond combined with a 0.33 arcsecond pixel scale (Barentsen et al., 2014) lead to a good ability to resolve sources even in crowded regions. In spite of this, IPHAS has a similar Q value as APASS, at 4.4×10^{-3}, indicating a similar relative level of contamination at the two catalogues' respective median magnitudes. This results in a contamination fraction of 10-15% at the faint end of the survey. Its much smaller PSF radius compared with APASS allows for a deeper survey at the same contamination level, or reduced contamination level at the same magnitude, as shown in Section 2.8.2.1.

While the survey does not provide astrometric uncertainties for individual stellar sources, the high quality of the photometry means that there is good agreement between empirical distribution uncertainties and astrometric uncertainties calculated as the image

scale length divided by the photometric SNR, as per King (1983).

2.8.2.1 APASS vs IPHAS

As was seen in Sections 2.8.1 and 2.8.2, both optical catalogues have a similar Q value – that is, the number of stars in an area the size of their PSF FWHM is similar. With the overlap in sky coverage and photometric bands, the separations of stars in common to both APASS and IPHAS with *Gaia* can be directly compared.

After matching both datasets to *Gaia* for $120 \leq l \leq 125$, $0 \leq b \leq 5$, IPHAS and APASS stars which matched to the same *Gaia* object were assumed to be themselves the same object. Stars were then selected with APASS astrometric uncertainties less than 0.15 arcsecond. Their separation distributions were then compared, as shown in Figure 2.7.

The theoretical AUF, a Rayleigh distribution with uncertainty 0.07 arcsecond (the typical positional uncertainty of the given subset of sources) matches the IPHAS distribution relatively well, with a small wing on the order of several percent, consistent with density contamination arguments. However, those same stars' positions are much more uncertain in APASS, caused in part by the differences in SNR, sky conditions etc., but additional broadening is caused by the vastly increased area subtended by stars on the sky in the APASS system.

As a consequence, the magnitude at which a given catalogue will reach approximately 33% contamination within its PSF can be compared separately. This value highlights the differences between APASS and IPHAS. The magnitude at which contamination of APASS sources up to five magnitudes fainter reaches 33% is $B \simeq 18.2$, which is approximately at the completeness limit of the survey in the uncrowded Galactic pole. However, the magnitude at which IPHAS suffers 33% five magnitude fainter contamination is $r = 23.4$, a few magnitudes fainter than its limiting magnitude of 20-21. This highlights the importance of spatial resolution on the contamination levels of photometric observations.

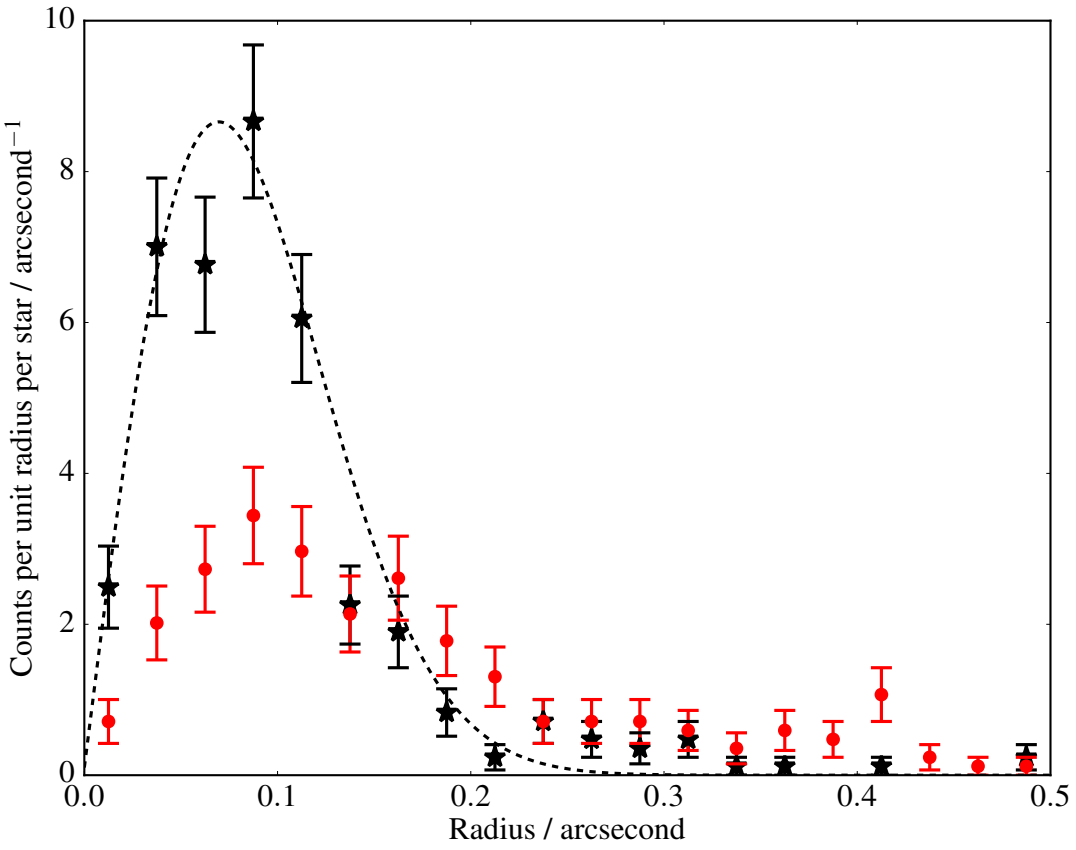


Figure 2.7: The effects of PSF resolution on the distribution of separations. Both IPHAS and APASS were matched to *Gaia* and those in common were plotted for $\sigma_{\mathrm{APASS}} < 0.15$ arcsecond, for IPHAS in black stars and APASS in red circles. The $\simeq$ 1 arcsecond FWHM of the IPHAS PSF gives contamination on the order of $\simeq$ 5% at an average of 15$^{\mathrm{th}}$ magnitude, whereas the 15 – 20 arcsecond aperture used for APASS leads to much increased contamination, causing a much broadened distribution. Theoretical distribution of separations is shown as a dotted line for reference.

2.8.3 2MASS

2MASS is frequently used to define the reference sky positions of those catalogues that came after it due to its all-sky completeness level (*WISE* and IPHAS both use it, for example), and therefore it is very important to understand the contamination levels that it suffers. However, it has a reasonably large PSF (FWHM $\simeq$ 2.5 arcsecond) and is a relatively faint ($K_s \simeq$ 16-17) survey. The contamination level rapidly increases with increasing magnitude and there are $\gtrsim 0.8$ stars in every 2MASS PSF at its limiting magnitude in the Galactic plane. This results in $Q = 1.3 \times 10^{-2}$, or one in three contaminated stars with contaminants up to five magnitudes fainter at $J = 13.4$.

The quoted uncertainties match the core region of the distribution to within 10%.

2.8.4 *WISE*

With its large, 10 arcsecond PSF and high SNR leading to faint limiting magnitudes, *WISE* is especially susceptible to crowding, leading to, on average, one faint star inside every PSF of stars with $W1 \simeq 13$. *WISE* has an especially large Q value, $\simeq 6 \times 10^{-2}$. Its 33% contamination level at the 1% flux level is reached at a very bright magnitude as well, with one in three stars of $W1 = 9$ suffering from a star $9 \leq W1 \leq 14$ inside its PSF.

In the Galactic plane, the core uncertainty is found to be twice the quoted uncertainty, explained by the large fraction of contaminated stars. However, the Galactic poles suffer much less from contamination with its reduced stellar density. It is found that at $\sigma_{\text{quoted}} \gtrsim$ 0.15 arcsecond the quoted uncertainties fit the distributions with only minor broadening. Core uncertainties are only 10-15% larger at these larger uncertainties, but below 0.15 arcsecond the core uncertainty plateaus requiring a constant σ to explain these brightest objects.

2.8.5 *Gaia*

As a survey dedicated to astrometry, *Gaia* has unparalleled precision in the positions of stars. Its PSF of 0.1 arcsecond FWHM leads to a very small Q value of 7.9×10^{-5}, 50 times

better than any other catalogue used, or a limiting magnitude contamination of $\simeq 0.1\%$. The magnitude at which contamination from stars five magnitudes fainter reaches 33% is $G = 30.7$, far fainter than the completeness limit of the survey. From this I am confident in using *Gaia* as the reference catalogue for quantifying the effects of contamination.

2.9 How to Deal with Contaminated Astrometric Detections

While the effect of unresolved objects inside stellar PSFs causing large wings to the probability distributions is explicit qualitatively, it is much more difficult to utilise it quantitatively. However, there are several ways to improve the matching process, depending on the specific requirements of the final catalogue of matches.

Two extremes of catalogue matching are the case where sources trusted to not be contaminated or be false positives must be the only ones returned, and the case where it is not necessarily an issue whether any sources are contaminated, and are also willing to accept a large number of false positives. The decision may also be motivated by whether it is acceptable that matches have detections with fluxes that are compromised by a second star in their PSFs in one or both of the respective catalogues.

In either case, it should be noted that there will be some situations, such as with *Gaia*-*WISE* matches, where one catalogue has a large PSF and the other has good spatial resolution, which will lead to a significant number of missed matches. These will be matches where one star contains within it as contamination a second object which is a separate entry in the opposite catalogue, which will lead to confusion in interpreting any results obtained. This will suggest that the faint *Gaia* source has a corresponding *WISE* magnitude below *WISE*'s completeness limit, which may not be the case in reality.

I also stress again that the contamination levels quoted here are upper limits, as active and passive deblending can help to resolve out overlapping objects, but note that this does not remove the effect entirely, as seen in the crowding out of stars (Figure 2.3).

2.9.1 Non-Contaminated Matches

First, when the goal is to only match those stars which are definitely true matches, but now additionally are not significantly flux contaminated, it is advisable to cut nearest neighbour matches at a minimum of $3\sigma_{\text{core}}$. Equivalently, σ_{core} should be used as the uncertainty in the AUF when considering probability-based matches.

I recommend examining sample distributions of nearest neighbour-matched separations. These should then be compared to their quoted uncertainty. If the quoted uncertainties are a good match to the empirical AUFs then use σ_{quoted}, but otherwise make empirical corrections to fit the slightly broadened distributions to match as required.

This will mostly capture the "clean" population, but will also increase the number of non-matches, as the AUF will not be sampling the extended tails of the contamination. This will potentially lead to the belief that the star was not detected in the opposing catalogue, with a cutoff radius that omits a large fraction of true matches. It will also still include some fraction of sources which are photometrically compromised, especially towards the fainter end of a given survey.

2.9.2 Full Coverage Matches

The other extreme is the case where the goal is to achieve a large catalogue with as many matches as possible, in which the effect of false positives or contaminated fluxes is unimportant.

In this case, the cutoff radius for a traditional nearest neighbour match should be some multiple of the largest PSF FWHM between the two catalogues, typically 1.5-2 FWHMs. Alternatively, if a probability-based matching system is being used, then it is advisable to construct a set of empirical AUFs for each astrometric uncertainty slice in turn, which will include the wings of the distributions. I will discuss this in more detail in Chapter 4.

These empirical functions are then used in place of f as described by Sutherland and Saunders (1992), g as per Naylor, Broos, and Feigelson (2013), Q_{χ^2} in Pineau et al.

(2017), LR_i of Rutledge et al. (2000), etc. These will increase the effective size of the area over which you can match between the catalogues, but will in turn increase the false match probability. Care should be taken when substituting any empirical functions into these probability-based matching methods, however, as any assumptions involving the use of Gaussian statistics (e.g., convolutions, mean positions, etc.) will no longer hold.

The *Gaia*-*WISE* case can be taken to demonstrate the effects of an empirical AUF. To do so, I matched the two catalogues using a probability-based matching process (see Chapter 3 for more details). The matching was done twice for two different astrometric PDFs. First, the AUFs used were purely Gaussian-based using σ_{quoted}, and second, the *WISE* AUF was empirically constructed. When comparing the number of returned cross-matches, the Gaussian-based AUFs returned approximately half the pairs that the empirically constructed AUFs matched (see Chapter 4 for a more in-depth discussion). Therefore, in crowded regions where the contamination of sources is high, probability-based matching using Gaussian statistics could result in as many as one in two true (albeit contaminated) counterparts being rejected as uncorrelated field objects.

2.10 Conclusions

I have presented an analysis of the distribution of *WISE* object positions with relation to *Gaia* positions to determine their AUF, the probability density function of a catalogue's detected positions as a function of distance. I have found that the core of the distribution of separations can be fit with Gaussian statistics, although they require broadening, which I fit for empirically. However, there is an additional, significant, non-Gaussian tail to the distributions which is explained by flux contamination from fainter stars lying undetected within the PSF of the brighter star. In addition, I have discussed the contamination levels of APASS, IPHAS, and 2MASS.

I have focussed on *WISE* in this chapter, as it is especially affected by this problem, because it reaches reasonably faint magnitudes in the infra-red and has a large PSF. However, it remains a problem for all catalogues, being an especially important consideration

for the next generation of very deep ground-based surveys, such as LSST, with its predicted depth in the optical of $r \simeq 25$ resulting in a theoretical Q value of approximately 4×10^{-2}. This means that at fainter magnitudes most detected objects will be contaminated by one or more faint objects in their PSF. In comparion, *Gaia* has a contamination level on the order of 0.1%, due to its 0.1 arcsecond FWHM PSF, meaning its positions should be robust against contamination.

Chapter 3

Improving Catalogue Matching By Supplementing Astrometry with Additional Photometric Information

A philosopher once asked, "Are we human because we gaze at the stars or do we gaze at the stars because we are human?" Pointless really. Do the stars gaze back? Now that's a question.

— Narrator, Stardust (2007)

3.1 Introduction

As surveys probe increasingly fainter magnitudes, leading in turn to a correspondingly fainter saturation magnitude, the effects of matching two catalogues with significantly differing dynamical ranges is rapidly becoming an issue. If two cleaned catalogues were matched, one might contain a faint detection but have removed a bright object due to saturation effects, while the other might contain the bright object as a good detection but have the faint object below its sensitivity limit. If these two objects were within a given critical match separation, it could appear that two incompatible objects were nearest neighbours to one another, which would result in an unphysical object in the merged dataset.

As discussed in Section 1.5, this crude nearest neighbour catalogue matching process can be improved with the use of the astrometric information each detection provides. This leads to a better description of the pairing of sources between two catalogues, as it is then linked to the certainty to which the observations can be known, changing the effective matching radius. There are three main approaches used in the literature to improve upon the nearest neighbour cross-match: the likelihood ratio, first used in this context in 1975 (Richter, 1975); the reliability (Sutherland and Saunders, 1992); and the extension to the cross-matching process using Bayes factors – the ratio of the probabilities of some dataset given two competing hypotheses – and including theoretical astrophysical models to describe the photometric likelihoods (Budavári and Szalay, 2008).

The likelihood ratio began with a purely astrometric consideration, the ratio of the probability that two sources on the sky were two detections of the same source to the probability that they were spuriously near to one another. de Ruiter, Willis, and Arp (1977) therefore define it as

$$LR(r) = dp(r|id)/dp(r|c) = \frac{1}{2\lambda} \exp\left(\frac{r^2}{2}(2\lambda - 1)\right), \tag{3.1}$$

with r the Mahalanobis distance between the two sources in question, and $\lambda = \pi\sigma_\alpha\sigma_\delta\rho(b)$ the equivalent dimensionless source density, the multiple of the error box area and sky density ρ, itself a function of Galactic latitude. More recently, the likelihood ratio has been extended to include the photometric information of the sources under consideration, usually given in the canonical form (e.g., Brusa et al., 2005)

$$LR = \frac{q(m)f(r)}{n(m)}, \tag{3.2}$$

where m is the magnitude of a given source in a chosen catalogue sources, and r the sky separation between the two sources. q is the probability density function (PDF) of the distribution of counterpart magnitudes, n is the surface density of background sources of magnitude m in the corresponding catalogue, and f is the PDF of the separation of the

sources given that they are counterparts, assumed to be Gaussian.

This form of the likelihood ratio was initially quoted by Sutherland and Saunders (1992); however, they then extend the formalism to include competing counterparts (i.e., the consideration that two sources in one catalogue might be astrometrically close to a source in a second catalogue) by defining the reliability

$$R_j = \frac{\Pr\left[S_j \cap \left(\bigcap_{k\neq j} U_k\right) \cap \left(\bigcap_{k'} E_{k'}\right)\right]}{\sum_i \Pr\left[S_i \cap \left(\bigcap_{k\neq i} U_k\right) \cap \left(\bigcap_{k'} E_{k'}\right)\right] + \Pr\left[(m_s > m_{\rm lim}) \cap \left(\bigcap_{k} U_k\right) \cap \left(\bigcap_{k'} E_{k'}\right)\right]}, \tag{3.3}$$

where S, U, and E are various events, that a given cell contains the correct source, contains a non-match source and is empty, respectively; i, j, and k are cell indices; and m_s and $m_{\rm lim}$ are the source and survey limiting magnitudes respectively. They then show that this form reduces to

$$R_j = \frac{L_j}{\sum_i L_i + (1 - Q)}, \tag{3.4}$$

where Q is a prior expected identification rate of counterparts. It is this form of the reliability that is used frequently in the literature (e.g., Fleuren et al., 2012), assuming a one-to-many catalogue cross-match, where the more crowded catalogue is defined by a constant cutoff magnitude $m_{\rm lim}$. Naylor, Broos, and Feigelson (2013) derive a more robust reliability formalism, allowing for a varying completeness limit – a more physical model, as the detection rate of the photometric catalogue is smooth with decreasing brightness, with no hard "cutoff" at a fixed magnitude. Their formalism is given as

$$P(i) = \frac{\frac{Xc(m_i)g(\Delta x_i,\Delta y_i)}{Nf(m_i)}}{1 - X + \sum_j \frac{Xc(m_j)g(\Delta x_j,\Delta y_j)}{Nf(m_j)}}, \tag{3.5}$$

intuitively similar to equation 3.4 with $Q = X$ and $L_i = Xc(m_i)g(\Delta x_i, \Delta y_i)/Nf(m_i)$, but the formalism effectively folds $m_s > m_{\rm lim}$ into c and f, no longer requiring an explicit catalogue-wide detection limit. They also provide the explicit probability of no match

($P(0)$), rather than simply quoting $1-Q$ (or $1-X$) as the chance, globally across the entire catalogue, of zero pair associations for a specific source. They also include a more robust treatment of the derivation of the counterpart magnitude distribution c – or q – accounting for the effect bright stars have on the detection of fainter nearby sources, improving upon the previous methodology of calculating $total(m)$ and subtracting a given background magnitude distribution to obtain $real(m)$ and then normalising to find q (as per, e.g., Fleuren et al., 2012).

Budavári and Szalay (2008), on the other hand, consider the competition between the hypothesis that n sources – across n catalogues – are n detections of one physical source, and the hypothesis that those n detections are of n distinct and separate objects. They begin by considering the ratio of the two hypotheses, often referred to as a Bayes factor,

$$B(H,K|D) = \frac{p(D|H)}{p(D|K)}, \tag{3.6}$$

where D is some dataset of the astrometric positions of the n detections, H is the hypothesis that the detections are all of the same source, and K is the hypothesis that the detections are of n different physical objects. They then marginalise over the true unknown position of each object, and, under the approximation that the spherical normal distribution is reducible to a Gaussian in the small angle approximation, and that the Gaussian is a good description of the uncertainty of the position of a given source, state the Bayes factor as

$$B = \frac{2}{\sigma_1^2+\sigma_2^2}\exp\left(-\frac{\psi^2}{2(\sigma_1^2+\sigma_2^2)}\right) \tag{3.7}$$

for the case of $n=2$ detections. They then construct the Bayes factor for the photometric detections, giving $B = B_{\rm pos}B_{\rm phot}$, where

$$B(H,K|D') = \frac{\int p(\boldsymbol{\eta}|H)\prod_{i=1}^{n} p_i(\boldsymbol{g}_i|\boldsymbol{\eta},H)\mathrm{d}^r\eta}{\prod_{i=1}^{n}\left[\int p(\boldsymbol{\eta}_i|K)p_i(\boldsymbol{g}_i|\boldsymbol{\eta},K)\mathrm{d}^r\eta_i\right]} \tag{3.8}$$

for observed fluxes $\boldsymbol{g}$ with physical properties $\boldsymbol{\eta}$ for a given model spectral energy distribution (SED) with theoretical fluxes in the given set of passbands used to detect the n

original measured positions.

However, in all methodology laid out in the literature the assumption that the astrometric probability is described by a Gaussian is still used. This does not correctly treat the effect of systematic astrometric perturbations. These effects include proper motion and the contamination from faint stars ("crowding", caused by the effects of finite pixel scale or point-spread-function width). I analysed these perturbations in the previous chapter, discussing the relative effect they have on matching separations. Additionally, no method to date simultaneously combines:

- the creation of magnitude relationships between catalogues without the use of prior astrophysical knowledge;
- photometric likelihoods which use these relationships bidirectionally, treating neither catalogue preferentially;
- a symmetric process which allows for the matching of equal astrometric precision datasets;
- the treatment of systematic effects in the astrometric detections of datasets;
- the consideration of all positionally correlated detections simultaneously in the resulting match probabilities;
- the explicit probability of a non-match of a star to any star in the opposing catalogue.

Here I will derive a matching process that is fully symmetric between the catalogues being matched, generalising Naylor, Broos, and Feigelson (2013), highlighting the assumptions that any asymmetric matching processes implicitly require. I will also discuss how to extend the matching process to multiple catalogues simultaneously, and briefly touch upon a few ways to reduce the complexity of such a matching process. I begin by introducing the problem and giving an overview of how to overcome it in Section 3.2. Section 3.3 gives a more rigorous derivation of the Bayesian formalism and the components of the equations. I then detail the forms that the astrometric, counterpart magnitude,

and unmatched star magnitude distributions take, in Sections 3.4 and 3.5. Section 3.6 gives two examples of the method applied to various catalogues. Section 3.7 then describes how to extend the method to three or more catalogues. Finally, I demonstrate consistency with previous asymmetric matching methods by showing how the equations presented here reduce back to the one-directional forms given by Naylor, Broos, and Feigelson (2013) in Section 3.8, giving concluding remarks in Section 3.9. Table 1.1 defines symbols used throughout.

3.2 Problem Setup

Before I formalise the problem, it is useful to show qualitatively how the method works. For this purpose, consider two catalogues that both contain detections in the same filter, with observations taken simultaneously with identical telescopes. One catalogue has good detections in the range $10 \leq m_\gamma \leq 16$, while the other catalogue has recorded sources with magnitudes $12 \leq m_\phi \leq 22$. There is a 100% counterpart rate in the dynamical range of both catalogues, $12 \leq m \leq 16$. The smallest non-trivial problem of matching between the two catalogues is the case where one star in catalogue γ and two stars in catalogue ϕ are positionally close to one another. All three stars are also sufficiently far away from all other stars that it can be assumed that no other star could be counterpart to any of the three of them. For illustration, let the given star in catalogue γ have a magnitude $m_\gamma = 14$. The two stars in catalogue ϕ are one bright star, $m_\phi = 14$, the correct counterpart, and a faint star, $m_\phi = 19$, that is slightly closer to the star in catalogue γ than its true counterpart. In this example, both stars in catalogue ϕ are close enough to be positionally likely to be matched with the star in catalogue γ. The two differing matches to the star in catalogue γ are my hypotheses: B, in the case of the bright object match, and F, for the case where the faint object is the counterpart.

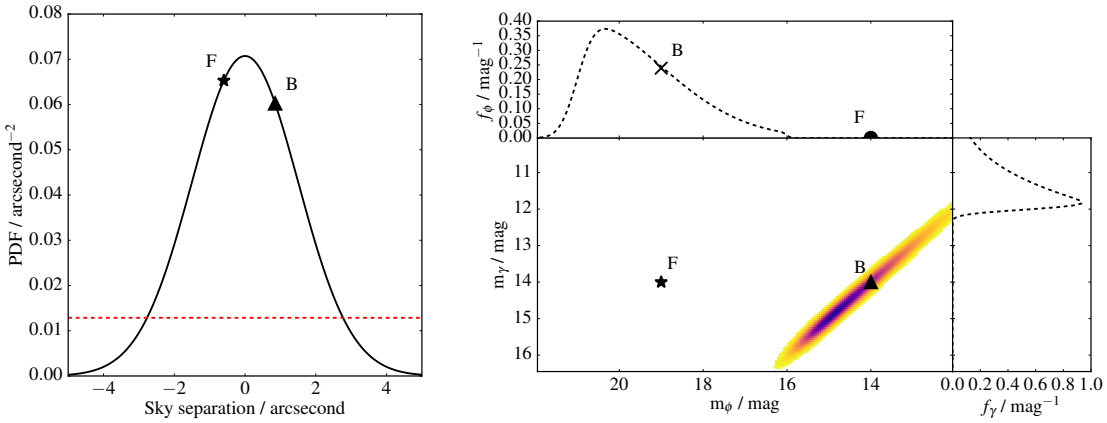
Figure 3.1a shows an example schematic for the probability of two stars being matched given their sky separation. As the distance between their measured positions increases, the probability of the two stars being counterparts to one another decreases until

they are more likely to be two unrelated stars. This is the point at which the counterpart PDF reaches the unmatched star probability density, indicated by the red dashed line. This probability is simply the chance of randomly placing unrelated stars in a small region of sky, based on the density of stellar sources nearby. If this PDF were used to match alone, the stars would simply be assigned as paired if their match probability is above the cut-off probability, or, equivalently, their separation is closer than the distance at which this transition occurs. In this case hypothesis F would be preferred, as the closest object to the star in catalogue γ is the fainter of the two catalogue ϕ stars.

If the knowledge of the relationship between magnitudes in both catalogues is introduced, an example of which is shown in Figure 3.1b, there is now a way to distinguish between the two sources in catalogue ϕ. If the intrinsic magnitude relationship between detections in each catalogue was known, the question could asked, based on the magnitude of two sources, whether they were likely to be the same star. Here both catalogues contain detections in the same filter, and therefore a detection in common between the two catalogues would measure the same brightness, to within experimental uncertainties.

Shown as dashed lines in the insets to Figure 3.1b are the probability densities of the objects in each catalogue (γ or ϕ) that do not have counterparts in the other catalogue (ϕ or γ). The unmatched PDF is the probability per unit magnitude that a star in catalogue γ, which does not have a corresponding entry in the catalogue ϕ, is measured at its given brightness. These are those stars that are either too bright, having saturated in the survey images, or are too faint, having too low a signal-to-noise ratio to be counted as a good detection, to be recorded in catalogue ϕ.

However, the probability of two stars being counterparts is a function of the brightness of both objects. This then leads to a two-dimensional function, an example of which is shown in the main panel of Figure 3.1b. In this example, using the same filters means that the likelihood is effectively a straight line along $y = x$ in magnitude-magnitude space, albeit blurred by observational uncertainties. This is also a PDF, this time per square magnitude, of a star having detected magnitudes m_γ and m_ϕ in the two catalogues respectively,



(a) The probability of a detection of a star in one catalogue being a given distance from its detection in a second catalogue. The solid line shows the probability density of two stars being counterparts as a function of their radial offset. The dashed line shows the constant probability density of unrelated stars. Any stars at a smaller sky separation than the distance at which the two lines are of equal probability (i.e., where the line of counterpart probability is higher than the line denoting the density of unrelated stars) would be assigned as counterparts to one another in a matching scheme. Triangle and star markers denote the separations of the matches in hypotheses B and F respectively.

(b) The probability distribution for counterpart stars in two catalogues as a function of magnitude. Inset figures shows the distribution of unmatched star magnitude probability for the two catalogues. In this case the surveys used the same photometric filter and therefore have a high counterpart probability of their magnitudes matching. The probability of being an unmatched star is high in the case where a star in catalogue γ is outside of the dynamical range of catalogue ϕ, and vice versa. Also marked are the probability densities for two hypotheses. Hypothesis B represents the case where two equal brightness ($m = 14$) objects have been assigned as counterparts (triangle, main figure) while a faint object ($m = 19$) in catalogue ϕ is unmatched (cross, inset figure). Hypothesis F represents the alternative match case, where the bright object in catalogue ϕ is unmatched (circle, inset figure), and the faint catalogue ϕ object is matched to the object in catalogue γ (star, main figure).

Figure 3.1: An example of star position and magnitude matching. Traditional matching would assign two detections as counterparts based purely on the positional probability, assigning the closest source only, preferring hypothesis F on astrometric arguments alone. However, the addition of the magnitude information allows for the correct matching of the true counterpart based on brightness, instead of simply positional correlation. The photometry allows for the pairing of the two objects with the magnitudes most likely drawn from an astrophysical object, accepting hypothesis B with the inclusion of the extra parameter space.

given that it is the same object detected twice.

For these hypotheses it is expedient to consider some shorthand notation. I denote the astrometric probabilities of two stars being drawn from a distribution of counterparts given their separation as $g(m_*, m_1)$, and of a star not having a counterpart as N. The photometric probability of two stars having their quoted magnitudes given that they are counterparts is $c(m_*, m_1)$, and the probability of a star having its magnitude given that it is not related to the other catalogue is $f(m_1)$. I also define the star in catalogue γ as m_*, the bright catalogue ϕ star as m_1, and the faint catalogue ϕ star as m_2.

Considering for the moment hypothesis B, a match between the star in catalogue γ and the bright catalogue ϕ star is required, while also not matching the faint catalogue ϕ star. This can be written as

$$P(B|m_*, m_1, m_2) = \frac{g(m_*, m_1)c(m_*, m_1)N_\phi f_\phi(m_2)}{O}, \tag{3.9}$$

where O is a normalisation, which will be discussed below. Alternatively, considering the opposite match,

$$P(F|m_*, m_1, m_2) = \frac{g(m_*, m_2)c(m_*, m_2)N_\phi f_\phi(m_1)}{O}. \tag{3.10}$$

The probability of the third case can also be expressed, in which neither star in catalogue ϕ is matched to the star in catalogue γ, as

$$P(C|m_*, m_1, m_2) = \frac{N_\gamma f_\gamma(m_*)N_\phi f_\phi(m_1)N_\phi f_\phi(m_2)}{O}. \tag{3.11}$$

In practice, this probability can be dismissed based on the assumption given previously that both catalogue ϕ stars are close enough to the catalogue γ object to be considered likely.

This means that $g(m_*, m_2) \gg N_\gamma N_\phi$. I include this third hypothesis for completeness, as the normalisation constant is simply the sum of the probability of all hypotheses, and thus

$$\begin{aligned} O =& N_\gamma f_\gamma(m_*) N_\phi f_\phi(m_1) N_\phi f_\phi(m_2) + g(m_*, m_1) c(m_*, m_1) N_\phi f_\phi(m_2) + \\ & g(m_*, m_2) c(m_*, m_2) N_\phi f_\phi(m_1). \end{aligned} \tag{3.12}$$

Considering the hypotheses B and F, their photometric probabilities can be focussed on, as the assumption has been made that both stars in catalogue ϕ are at roughly equal sky separation from the catalogue γ source, and thus $g(m_*, m_1) \simeq g(m_*, m_2)$.

Hypothesis F (the faint star being the counterpart) leads to a low photometric probability density for all stars, with a low counterpart likelihood $c(m_*, m_2)$, and low field likelihood $f(m_1)$. However, the opposite hypothesis, B (the bright star being the counterpart), has a high probability in both the photometric match between the two bright stars and the faint catalogue ϕ star being a field star. The main panel of Figure 3.1b shows the probability densities for the counterpart matches for both hypotheses. Here the likelihood of the bright catalogue γ object being the same object as the faint catalogue ϕ object photometrically is low, but the bright stars in both catalogues have a high probability of being the same source. Additionally, the hypotheses can be further differentiated on the probability of the unmatched object. Along a similar line of reasoning, the unmatched object probability densities in the top inset figure can be considered. The rejected faint catalogue ϕ star in hypothesis B has a high unmatched probability density, whereas hypothesis F leads to a low unmatched star probability density.

The combination of these two probability densities can be used, for any matched and, just as usefully, unmatched objects, to help break any degeneracies in the astrometric matches. Such cases, where stars may have similar Mahalanobis distances, would be difficult to resolve with just the astrometric probability. This is especially significant when the astrometric probability is much higher than the unrelated source density against which a non-match is to compared. The result in the example is that while the bright

catalogue ϕ object has a slightly larger sky separation (and would therefore not be matched astrometrically, by a nearest neighbour scheme or purely astrometric probability match; see Figure 3.1a for comparison of the objects' sky separations), it is overwhelmingly more favourable as the counterpart. The photometric information can be used to correctly select the bright counterpart over the faint interloper.

While I have focussed on the case where two stars are potential matches to a given object, the trivial case can also be considered. In this instance there is only one star from each catalogue, with the determination as to whether they are counterparts or unrelated objects in question. If the stars were within the cut-off radius of a traditional nearest neighbour-match method they would be paired automatically. However, the flexibility of the probability-based matching scheme allows for the direct comparison of the likelihood of the two stars being at their separations and magnitudes. Both the case where they are the same star observed in two catalogues and the case where they are two different unrelated objects can be examined before considering them as counterparts.

3.3 Constructing the Bayesian Framework

Each photometric catalogue can be considered to be a three-dimensional position-position-magnitude cube. Each small square of sky plane is either filled with an object's detection, or blank and thus a non-detection. However, each position-position square that contains a star only has a filled cell at the recorded stellar magnitude. When matching two of these catalogues together, the question is being asked whether a given filled cell in catalogue γ corresponds to a filled cell in catalogue ϕ, or if they are unrelated.

Following a similar notation to that of section 2.1 of Sutherland and Saunders (1992), I define a "cell" to be have a volume $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}m$. I also define various events for detections and non-detections of objects in these cells, given in Table 3.1. In terms of notation, for each event the subscript refers to the specific catalogue (in this case, either γ or ϕ), whereas the superscript refers to the individual cell (e.g., i or j) in the given catalogue.

Event	Notation
cell i is empty in catalogue γ	E^i_γ
cells i and j are occupied by a star that is in both catalogue γ and ϕ, respectively	$S^{ij}_{\gamma\phi}$
cell i is occupied by star in catalogue γ that is not in catalogue ϕ	U^i_γ

Table 3.1: Table showing the definitions of various events for catalogue matching.

3.3.1 The Match Hypotheses

Considering the case where one star in each catalogue is matched, and all other stars are unrelated, hypothesis H_a, an expression for the likelihood of the data given this hypothesis can also be written,

$$P(D|H_a) \propto P\left[S^{kl}_{\gamma\phi} \cap \left(\bigcap_{i\neq k} U^i_\gamma\right) \cap \left(\bigcap_{i'} E^{i'}_\gamma\right) \cap \left(\bigcap_{j\neq l} U^j_\phi\right) \cap \left(\bigcap_{j'} E^{j'}_\phi\right)\right]. \tag{3.13}$$

Here $S^{kl}_{\gamma\phi}$ is the probability that a given star occupies cell k in catalogue γ and cell l in catalogue ϕ, $E^{i'}_\gamma$ is the probability that cell i' in catalogue γ is empty, and U^i_γ is the probability that cell i is occupied by a star in catalogue γ which is not in any cells in catalogue ϕ. Equation 3.13 runs over k and l, the cells containing only matched stars; i and j, the cells filled with unrelated stars; and i' and j', the empty cells, for each catalogue respectively.

Now, considering the case where no stars are in common between the two catalogues, denoting it as H_0, a second hypothesis likelihood can be written as

$$P(D|H_0) \propto P\left[\left(\bigcap_{i} U^i_\gamma\right) \cap \left(\bigcap_{i'} E^{i'}_\gamma\right) \cap \left(\bigcap_{j} U^j_\phi\right) \cap \left(\bigcap_{j'} E^{j'}_\phi\right)\right], \tag{3.14}$$

where, again, i and j run over all filled cells and i' and j' run over all other cells.

At this point Bayes' rule can be applied to obtain hypothesis posteriors, given by

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}. \tag{3.15}$$

Here the evidence, $P(D)$, is simply the sum over all possible hypotheses; i.e., the sum over

the null hypothesis H_0 and all possible combinations of H_a,

$$P(D) = P(D|H_0)P(H_0) + \sum_a P(D|H_a)P(H_a). \tag{3.16}$$

This requires a choice of prior. As any combination of unmatched and matched objects must be accepted with equal probability, the hypothesis ensemble has an indifferent prior, and thus $P(H_0) = P(H_a)$ for all a. The prior can then simply be neglected from the combination of equations 3.15 and 3.16. In addition, the sum over i' and j' can be omitted, as all empty cells remain empty in all hypotheses and are assumed to be independent of filled cells and each other. The terms simply cancel in the numerator and denominator of equation 3.15. Thus my slightly modified version of equation 4 of Sutherland and Saunders (1992) is

$$P(H_a|D) = \frac{P\left[S^{kl}_{\gamma\phi} \cap \left(\bigcap_{i\neq k} U^i_\gamma\right) \cap \left(\bigcap_{j\neq l} U^j_\phi\right)\right]}{P\left[\left(\bigcap_i U^i_\gamma\right) \cap \left(\bigcap_j U^j_\phi\right)\right] + \sum_s \sum_t P\left[S^{st}_{\gamma\phi} \cap \left(\bigcap_{i\neq s} U^i_\gamma\right) \cap \left(\bigcap_{j\neq t} U^j_\phi\right)\right]}. \tag{3.17}$$

The independent cell assumption can be extended and thus the probabilities split. Therefore equation 3.17 becomes

$$P(H_a|D) = \frac{P\left(S^{kl}_{\gamma\phi}\right) \prod_{i\neq k} P\left(U^i_\gamma\right) \prod_{j\neq l} P\left(U^j_\phi\right)}{\prod_i P\left(U^i_\gamma\right) \prod_j P\left(U^j_\phi\right) + \sum_s \sum_t P\left(S^{st}_{\gamma\phi}\right) \prod_{i\neq s} P\left(U^i_\gamma\right) \prod_{j\neq t} P\left(U^j_\phi\right)}, \tag{3.18}$$

with the additional equation

$$P(H_0|D) = \frac{\prod_i P\left(U^i_\gamma\right) \prod_j P\left(U^j_\phi\right)}{\prod_i P\left(U^i_\gamma\right) \prod_j P\left(U^j_\phi\right) + \sum_s \sum_t P\left(S^{st}_{\gamma\phi}\right) \prod_{i\neq s} P\left(U^i_\gamma\right) \prod_{j\neq t} P\left(U^j_\phi\right)}. \tag{3.19}$$

Here $P(H_a|D)$ is a stand-in for R_j, the reliability of an object (Sutherland and Saunders,

1992), and I include the extra probability $P(H_0|D)$, introduced by Naylor, Broos, and Feigelson (2013). However, it is important to note that only *unrelated* cells are independent, and therefore the probabilities of a match between the two catalogues cannot be separated, and so $S^{st}_{\gamma\phi}$ must still be considered jointly.

3.3.2 Event Probabilities

Forms for each event are required, for which I follow the notation of Naylor, Broos, and Feigelson (2013). Position and magnitude are also assumed to be independent, and therefore are separable. Event U, the probability of an unrelated cell, is written

$$P\left(U^i_\gamma\right) = N_\gamma \, \mathrm{d}x \, \mathrm{d}y \, f_\gamma(m_i) \, \mathrm{d}m, \tag{3.20}$$

where the probability of an unmatched star being in a given position is simply N_γ, the number density of unmatched stars, multiplied by $\mathrm{d}x\mathrm{d}y$, the cell sky area. Additionally, the probability of an unmatched star having magnitude m to $m + \mathrm{d}m$ is $f_\gamma(m_i)$, the unmatched star magnitude distribution at m_i, multiplied by $\mathrm{d}m$.

The function for the probability of two stars matching between the two catalogues is slightly more involved. These require joint probabilities, which are written as

$$P\left(S^{kl}_{\gamma\phi}\right) = g(x_k, y_k, x_l, y_l) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}x \, \mathrm{d}y \, c(m_k, m_l) \, \mathrm{d}m \, \mathrm{d}m \tag{3.21}$$

for now, with each term being expanded separately. Here g is the probability density, per degree[4], of two stars being counterparts to the same object with their recorded sky positions, while c is the probability density, per square magnitude, that an object has its given quoted magnitudes in both catalogues.

No matter what combination of stars are in question, the same volume $(\mathrm{d}x)^2(\mathrm{d}y)^2(\mathrm{d}m)^2$ is considered for all stars. Therefore the volume terms in equations 3.20 and 3.21 cancel, and I make the change from pure probability to probability densities and a change from P

to p in my notation.

3.3.2.1 Astrometric Match Probability Density Function

The probability that the stars are counterparts requires the probability that star k and l are drawn from the same original sky position. This can be found by deriving the probability that the stars both originated from the same, but unknown, sky position x_0, y_0. It is relatively straightforward to compute the probability of two different detections of an object being at two sky positions given a known "true" position. However, it is more involved to obtain the probability of the two objects originating from the same position without prior knowledge. Handling this issue in a Bayesian fashion, all "true" positions can be marginalised over, giving

$$\begin{aligned} g(x_k, y_k, x_l, y_l) &= \int_{-\infty}^{+\infty}\!\!\!\int p(x_k, y_k, x_l, y_l|x_0, y_0)p(x_0, y_0)\,\mathrm{d}x_0\,\mathrm{d}y_0 \\ &= \int_{-\infty}^{+\infty}\!\!\!\int h_\gamma(x_0 - x_k, y_0 - y_k)h_\phi(x_l - x_0, y_l - y_0)p(x_0, y_0)\,\mathrm{d}x_0\,\mathrm{d}y_0, \end{aligned} \tag{3.22}$$

where h_γ and h_ϕ are the rotationally symmetric (i.e., $f(x,\, y) = f(-x,\, -y)$) distributions of the astrometric uncertainties for catalogues γ and ϕ respectively. I assign a flat prior on x_0 and y_0,

$$p(x_0, y_0) = N_\mathrm{c}, \tag{3.23}$$

the number of objects in common between the two catalogues per unit area. The details of how I calculate this number are described in Section 3.6.3. Equation 3.23 can be substituted into equation 3.22 obtaining

$$g(x_k, y_k, x_l, y_l) = N_\mathrm{c} \int_{-\infty}^{+\infty}\!\!\!\int h_\gamma(x_0 - x_k, y_0 - y_k) \times h_\phi(x_l - x_0, y_l - y_0)\,\mathrm{d}x_0\,\mathrm{d}y_0. \tag{3.24}$$

The terms $\Delta x_{kl} = x_l - x_k$ and $\Delta y_{kl} = y_l - y_k$ can also be substituted, giving

$$g(x_k, y_k, x_l, y_l) = N_c \iint_{-\infty}^{+\infty} h_\gamma(x_0 - x_l + \Delta x_{kl}, y_0 - y_l + \Delta y_{kl}) \times h_\phi(x_l - x_0, y_l - y_0)\, \mathrm{d}x_0\, \mathrm{d}y_0. \tag{3.25}$$

Substituting $x = x_l - x_0$ and $y = y_l - y_0$ obtains

$$\begin{aligned} g(x_k, y_k, x_l, y_l) &= N_c \iint_{-\infty}^{+\infty} h_\gamma(\Delta x_{kl} - x, \Delta y_{kl} - y) h_\phi(x, y)\, \mathrm{d}x\, \mathrm{d}y \\ &= N_c \times (h_\gamma * h_\phi)(\Delta x_{kl}, \Delta y_{kl}). \end{aligned} \tag{3.26}$$

Here $(h_\gamma * h_\phi)(\Delta x_{kl}, \Delta y_{kl})$ denotes the convolution of the functions h_γ and h_ϕ, measured at position Δx_{kl}, Δy_{kl}. To streamline my notation, I redefine equation 3.26 to be

$$g(x_k, y_k, x_l, y_l) = N_c G(\Delta x_{kl}, \Delta y_{kl}). \tag{3.27}$$

The resulting distribution is then a convolution of the two catalogues' individual astrometric uncertainty functions (AUFs; Section 1.6), multipled by a prior term. This result is often quoted by other authors for the specific case where G is Gaussian in both catalogues (e.g., equation 16 of Budavári and Szalay, 2008). In this simple case the convolution of the two functions is itself a Gaussian with uncertainty $\sigma^2_{\mathrm{new}} = \sigma^2_k + \sigma^2_l$, evaluated at Δx_{kl}, Δy_{kl}. However, I know of no formal proof in the general case, although I note similarities between my equation 3.22 and equation 9 of Budavári and Szalay (2008) and, albeit without the prior term, equation 38 of Pineau et al. (2017).

It should be noted that it cannot be assumed *a priori* that G will be a Gaussian, as the individual catalogue AUFs cannot themselves be assumed Gaussian. This is due to systematic effects such as proper motions, or the effects of faint contaminants within detected stars' point-spread functions (PSFs) on their measured positions (see Chapter 2

for more details). My more general formalism allows for the inclusion of the treatment of such systematics (see Chapter 4 for a discussion of the effect this treatment has on the matching in highly contaminated crowded fields). Additionally, I note that this proof is only true for the specific case of matching two catalogues; see Section 3.7 for the more general treatment of 3 or more catalogues.

3.3.2.2 Photometric Match Probability Density Function

The probability of two stars being related as a function of their respective magnitudes is also required. If information about the intrinsic relationship between sources in both catalogues was available, the stars' unknown "true" stellar magnitudes could be marginalised over. This would be analogous to equation 3.22, and give

$$\begin{aligned} c(m_k, m_l) &= \int_{-\infty}^{+\infty}\int p(m_k, m_l|m_a, m_b)p(m_a, m_b)\,\mathrm{d}m_a\,\mathrm{d}m_b \\ &= \int_{-\infty}^{+\infty}\int p(m_k|m_a)p(m_l|m_b)p(m_a, m_b)\,\mathrm{d}m_a\,\mathrm{d}m_b. \end{aligned} \tag{3.28}$$

The likelihoods in this case would be

$$p(m_k|m_a) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(\frac{-(m_k - m_a)^2}{2\sigma_k^2}\right) \tag{3.29}$$

and

$$p(m_l|m_b) = \frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(\frac{-(m_l - m_b)^2}{2\sigma_l^2}\right), \tag{3.30}$$

and $p(m_a, m_b)$ would represent the prior, intrinsic joint magnitude distribution on counterpart magnitudes m_a and m_b.

In practice, however, it is not possible to disentangle the observational uncertainties $\big(p(m_k|m_a), p(m_l|m_a)\big)$ and intrinsic relationships $\big(p(m_a, m_b)\big)$ from the data which measure $c(m_k, m_l)$, and therefore c is measured directly. However, I include this description for symmetry and completeness.

3.3.3 Combined Bayesian Probabilities

3.3.3.1 One Match Equation Form

For compact notation in this subsection, I define the following terms:

$$G(\Delta x_{kl}, \Delta y_{kl}) = G^{kl}_{\gamma\phi},\ c(m_k, m_l) = c^{kl}_{\gamma\phi},\ f_\gamma(m_i) = f^i_\gamma. \tag{3.31}$$

This notation follows a similar style to that previously, where each PDF (G, c, and f) has a subscript denoting which catalogue it refers to, and a superscript which identifies the star in the given catalogue. My revised probabilities for H_0 and H_a are therefore

$$P(H_a|D) = \frac{N_c G^{kl}_{\gamma\phi} c^{kl}_{\gamma\phi} \prod\limits_{i\neq k} N_\gamma f^i_\gamma \prod\limits_{j\neq l} N_\phi f^j_\phi}{\prod\limits_i N_\gamma f^i_\gamma \prod\limits_j N_\phi f^j_\phi + \sum\limits_s \sum\limits_t N_c G^{st}_{\gamma\phi} c^{st}_{\gamma\phi} \prod\limits_{i\neq s} N_\gamma f^i_\gamma \prod\limits_{j\neq t} N_\phi f^j_\phi}, \tag{3.32}$$

and

$$P(H_0|D) = \frac{\prod\limits_i N_\gamma f^i_\gamma \prod\limits_j N_\phi f^j_\phi}{\prod\limits_i N_\gamma f^i_\gamma \prod\limits_j N_\phi f^j_\phi + \sum\limits_s \sum\limits_t N_c G^{st}_{\gamma\phi} c^{st}_{\gamma\phi} \prod\limits_{i\neq s} N_\gamma f^i_\gamma \prod\limits_{j\neq t} N_\phi f^j_\phi}. \tag{3.33}$$

Equations 3.32 and 3.33 represent the fundamental result of this section, being generalised versions of previous formulations (e.g., equation 4 of Sutherland and Saunders, 1992 or equation 7 of Naylor, Broos, and Feigelson, 2013). These equations give the probability of one star in catalogue γ and one star in catalogue ϕ being counterparts, or the probability of there being no counterpart between stars in catalogues γ and ϕ, respectively.

However, this formulation is limited, as shown by some simple example catalogues. Consider the case where both catalogue γ and catalogue ϕ contain two objects each – $\gamma_{1,2}$ and $\phi_{1,2}$ respectively. The formulation used by Naylor, Broos, and Feigelson (2013) assumes that each X-ray source (catalogue γ object) does not compete with any other X-ray source for potential IR detection counterparts (catalogue ϕ objects). In such a case, equation 7 of Naylor, Broos, and Feigelson (2013) would have two potential counterpart pairings, $\gamma_1\phi_1$ and $\gamma_1\phi_2$, as γ_2 is assumed to not be positionally close to these catalogue ϕ objects. In equation 3.32 this assumption has been lifted, allowing for two more hypotheses: $\gamma_2\phi_1$ and $\gamma_2\phi_2$, the pairing of the second catalogue γ object with either catalogue ϕ object.

However, equation 3.32 assumes that, no matter how many stars are detected in either catalogue, at most one star was detected twice, and therefore only one star from one catalogue is a counterpart to one star in the other catalogue. This might be useful in many situations, where one catalogue is so sparse that two sources cannot possibly "compete" for the same source in the opposing catalogue (e.g., Naylor, Broos, and Feigelson, 2013), but is not necessarily the case in general. In crowded Galactic plane regions, for example, there may be a scenario where the recorded positions of multiple stars from each catalogue cannot be disentangled. It might be reasonable to assume that most of the objects recorded in both catalogues are the same objects detected twice. In this scenario, my example catalogues would have two additional hypotheses that must be included: $\gamma_1\phi_1$ and $\gamma_2\phi_2$; and $\gamma_1\phi_2$ and $\gamma_2\phi_1$.

It is then no longer possible to make the assumption that there are either zero or one multiply detected object, as has been made throughout Section 3.3 thus far. To account for the cases where the assigning of more than one counterpart pairing is required it must be possible to express equations 3.32 and 3.33 in a more general form.

3.3.3.2 Multiple Match Equation Form

To account for multiple star pairings, equations 3.32 and 3.33 can be extended to any permutations of potential pairings between the catalogues γ and ϕ. For a given hypothesis, the probability that there are k matches between the two catalogues is to be calculated. Here ζ is a given k-permutation of catalogue γ, and λ is a given k-combination of catalogue ϕ. The use of permutations of one catalogue and combinations of the second catalogue avoids the repeated consideration of the same hypothesis – pairing A with B and C with D is the same as matching C with D and A with B.

For example, if there are two matching stars between γ and ϕ then $k = 2$. If there the four stars in γ, then $\gamma = \{1, 2, 3, 4\}$, for instance. In this case, one potential subset of counterparts could be $\zeta = \{2, 4\}$. The probability that all stars which have been "paired" match, and all other stars are unmatched in both catalogues, is required. H_0 is then the hypothesis that $k = 0$, and H_a is the hypothesis that there is one matched star in ζ, paired with the star in λ.

My full equation is

$$P(\zeta, \lambda, k|\gamma, \phi) = K \times \prod_{\delta \notin \zeta \cap \delta \in \gamma} N_\gamma f_\gamma^\delta \prod_{\omega \notin \lambda \cap \omega \in \phi} N_\phi f_\phi^\omega \prod_{i=1}^{k} N_c G_{\gamma\phi}^{\zeta_i \lambda_i} c_{\gamma\phi}^{\zeta_i \lambda_i}, \tag{3.34}$$

where K is a normalisation constant, which can generally be expressed as the sum of the posterior probability of no matches plus the summation over all possible match number permutations. The normalisation requires a sum over three indices. First, the number of matches, k, from 0 to the number of objects in the smallest catalogue, resulting in a 100% match rate, $\min(n_\gamma, n_\phi)$. Second, each of the k-permutations of γ, the set of which I define as Γ_k. Finally, the normalisation must sum over each of the k-combinations of ϕ, the set of which is Φ_k. Thus

$$K = \sum_{k=0}^{\min(n_\gamma, n_\phi)} \sum_{\zeta \in \Gamma_k} \sum_{\lambda \in \Phi_k} \prod_{\delta \notin \zeta \cap \delta \in \gamma} N_\gamma f_\gamma^\delta \prod_{\omega \notin \lambda \cap \omega \in \phi} N_\phi f_\phi^\omega \prod_{i=1}^{k} N_c G_{\gamma\phi}^{\zeta_i \lambda_i} c_{\gamma\phi}^{\zeta_i \lambda_i}. \tag{3.35}$$

While the equations presented are flexible in their application and set size, it is impractical to consider the entire dataset as one entity. I therefore limit the set size to those stars positionally close to another star in the set. This limitation results in a large number of star "islands". These islands could potentially reduce to the situation considered initially, with one star in one catalogue having multiple potential counterparts, for which equations 3.32 and 3.33 would be applicable. Typical number of stellar overlaps are ≤5, with the majority of stars only overlapped by 1-3 objects in the catalogue they are being matched to. The complexity can therefore be reduced in most cases back to that seen in equations 3.32 and 3.33. In more complicated island permutations, with multiple stars in each catalogue under consideration, the more general equations 3.34 and 3.35 should be used.

In the next two sections I will expand my terms for G, c, and f, and detail how to calculate them.

3.4 Functional Forms of Astrometric Distributions

The astrometric PDF G is defined for the two catalogue match as the convolution of the AUFs of the two stars in question (see Section 3.3.2.1). As such, functions for the AUFs are required. For the rest of this chapter I will assume that the probability of detecting a star with a given uncertainty, at a given offset $(x,\ y)$ from its implied true origin, is given by a Gaussian. These AUFs describe how accurately the position of the star is known, which is vital for this probabilistic matching process.

It can be shown, as I did in Chapter 2, that the empirical AUFs of a given catalogue may not be purely Gaussian, but are best described as broadened core distributions and large, non-Gaussian wings. These effects are caused by systematics such as proper motion or contamination from unresolved, faint objects inside the PSF of the bright star. However, for the purposes of testing this method in Section 3.6 I will focus on photometric catalogues with sufficiently small PSFs and number densities such that the average number of stars per PSF is low, which will limit the effect of the contamination to a few percent of stars at

most.

In general, the AUFs can be two-dimensional elliptical Gaussians, meaning uncertainties in the orthogonal α (or right ascension), and δ (or declination) directions are required, as well as the correlation between the two, ρ. The transformations from semi-major axis a, semi-minor axis b, and position angle east of north θ, if required, are given by

$$\begin{gathered}\sigma_\alpha = \sqrt{a^2 \sin^2(\theta) + b^2 \cos^2(\theta)},\ \sigma_\delta = \sqrt{a^2 \cos^2(\theta) + b^2 \sin^2(\theta)}, \\ \rho = \frac{(a^2 - b^2)\sin(\theta)\cos(\theta)}{\sigma_\alpha \sigma_\delta}.\end{gathered} \tag{3.36}$$

For a two-dimensional PDF centered at the origin with covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\delta \\ \rho\sigma_\alpha\sigma_\delta & \sigma_\delta^2 \end{pmatrix}. \tag{3.37}$$

The formulation of a given Gaussian AUF is then

$$h(\Delta\alpha, \Delta\delta, \sigma_\alpha, \sigma_\delta, \rho) = \frac{\exp\left(-\frac{1}{2\sqrt{1-\rho^2}}\left(\frac{(\Delta\alpha)^2}{\sigma_\alpha^2} + \frac{(\Delta\delta)^2}{\sigma_\delta^2} - \frac{2\rho\Delta\alpha\Delta\delta}{\sigma_\alpha\sigma_\delta}\right)\right)}{2\pi\sigma_\alpha\sigma_\delta\sqrt{1-\rho^2}}. \tag{3.38}$$

Note that when dealing with offsets in right ascension, I include the cosine of the declination to convert the separations to seconds of arc.

In this case I am considering the matching of three catalogues: INT Photometric Hα Survey (IPHAS; Drew et al., 2005; Barentsen et al., 2014); Two Micron All Sky Survey (2MASS; Skrutskie et al., 2006); and *Gaia* Data Release 1 (DR1; Gaia Collaboration et al., 2016b; Gaia Collaboration et al., 2016a). For the rest of this chapter I shall assume that the 2MASS and *Gaia* astrometry are well modelled by Gaussians with uncertainties as quoted in their respective catalogues. IPHAS, however, does not quote individual source positional uncertainties, and I therefore use the relation given by King (1983),

$$\sigma_\alpha = \sigma_\delta = \sqrt{(0.05\,\text{arcsecond})^2 + \left(\frac{\text{FWHM}_{\text{IPHAS}}}{2\sqrt{2\log(2)} \times \text{SNR}_{\text{IPHAS}}}\right)^2} \tag{3.39}$$

where the full width at half maximum (FWHM) of the observational seeing is taken from the IPHAS catalogue for every star individually, and the signal-to-noise ratio (SNR) can be calculated from the statistical photometric uncertainty, also quoted individually for every star. The 0.05 arcsecond is the typical systematic astrometric uncertainty, the average plate solution residual, using 2MASS sources to constrain the absolute astrometric transformation of an image. I use this combined uncertainty as the standard deviation in the Gaussian AUFs for the IPHAS data.

The argument laid out by King (1983) follows from the consideration of $l\,f_i + \alpha b = n_i$, where l is the expected counts from a given star, f_i is the fraction of the profile in a pixel i, α is the area of said pixel, b is some background pixel counts, and n_i is the observed counts of the ith pixel. This equation, in essence, attempts to fit the data – the observed counts n – with a uniform background count level and a description of the PSF affecting the point source. Initially used to solve for the SNR of the source including the PSF profile and sky brightness, the inclusion of first-order differential corrections to the equation allows for the derivation of positional uncertainty. Combining these two equations – the original model fit to the observed pixel counts, and its first-order differential – results in a form similar to that given in equation 3.39, where the ratio of the statistical astrometric uncertainty to the characteristic length scale of the observation – σ in the case that the PSF is approximated as a Gaussian – is approximately equal to the ratio of the uncertainty on the flux to the flux, or the inverse of the SNR.

As G is the function to be calculated, the two Gaussian distributions must be convolved together. To do so, the given covariance matrices of the two functions (equation 3.37) are simply added together, giving a new σ_α, σ_δ, and ρ, then used in equation 3.38.

3.5 Functional Forms of Magnitude Distributions

Now that the probability of correlation between two objects positionally is found, the probability of their relatedness in magnitude space must be considered. In this case, two possibilities must be considered. First, that each object in catalogue γ is an unmatched object, unrelated to anything in catalogue ϕ. Second, the two objects have magnitudes that have high likelihoods of being the same object detected in both catalogues. For these two cases the counterpart probability density function must be built, which I denote as $c(m_i, m_j)$, and the unmatched ("field"; Naylor, Broos, and Feigelson, 2013) star PDFs $f(m_i)$ and $f(m_j)$. f is a PDF, the probability per unit magnitude of a star having its observed magnitude, given that it is unpaired (see, e.g., insets to Figure 3.1b). c is also a PDF, probability per unit γ magnitude per unit ϕ magnitude, of two objects having their respective magnitudes given the assumption that they are counterparts to one another (Figure 3.1b).

These functions are constructed from the catalogues in situ. The magnitudes of all stars in catalogue γ positionally unrelated to any star in catalogue ϕ must therefore be considered to build the unmatched magnitude distribution. Similarly the magnitudes of stars positionally close to one another must be considered to build the counterpart likelihood. This simplistic approach is shown graphically in Figure 3.2. The black line shows the distribution of IPHAS stars within 3 arcseconds of *Gaia* stars of magnitude $14 \leq G \leq 15$ in the region $120 \leq l \leq 125$, $0 \leq b \leq 5$. The similar passbands in the i and G filter result in most of these objects having $i \simeq 15$. The red line shows the distribution of i magnitudes of stars with no *Gaia* object within 3 arcseconds of their position. These global properties indicate that the "correct" match, on photometric grounds, to a *Gaia* star of $G = 15$ should have an i-band detection of roughly 15th magnitude as well, while an IPHAS star of $i = 18$ or fainter is unlikely to match a bright *Gaia* source.

The unmatched star distributions are fairly straightforward, requiring merely the omission of any stars within sufficiently large circles of stars in the other catalogue, the details of which are described in Section 3.6.3. f can be populated by simply recording

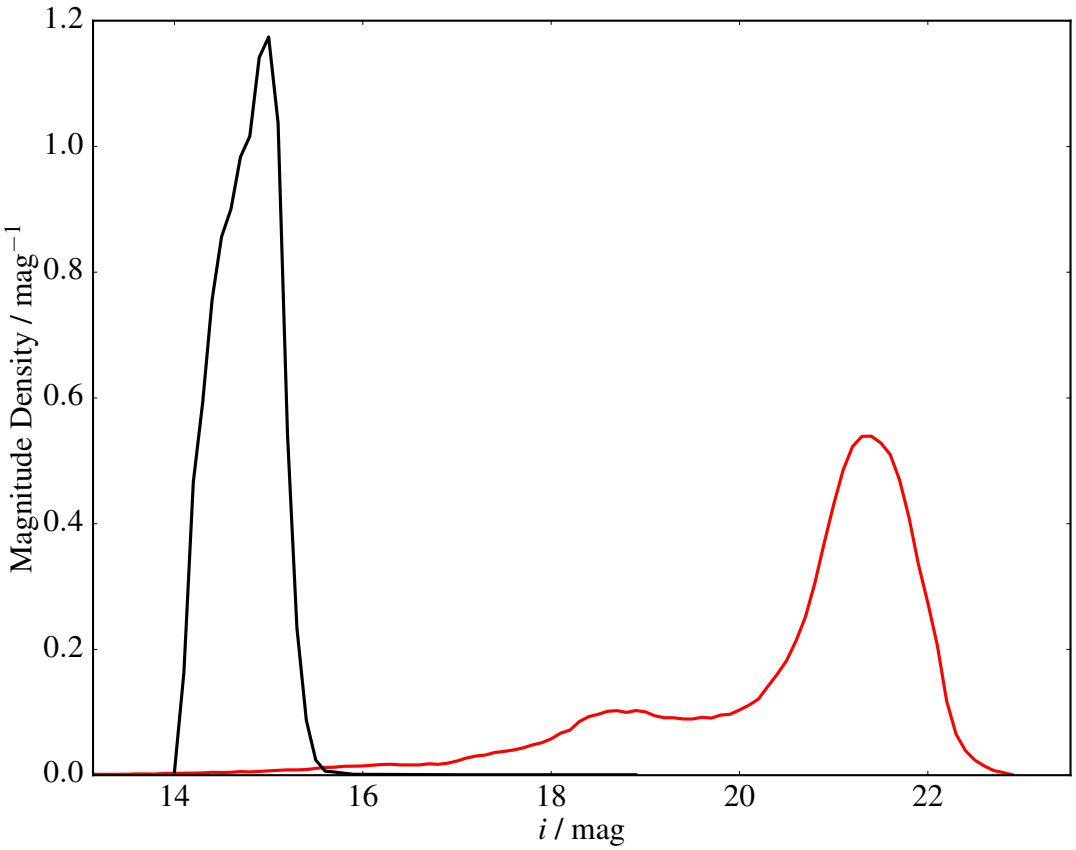


Figure 3.2: The distributions of matched and unmatched IPHAS stars. The black line shows the i magnitudes of all stars within 3 arcseconds of *Gaia* stars of $14 \leq G \leq 15$, while the red line shows the i magnitude of all IPHAS stars with zero *Gaia* stars within 3 arcseconds of their location. The vastly differing resulting PDFs highlight the power of the addition of the magnitude information: it is very likely that a *Gaia* star with $G = 15$ would match an IPHAS star with $i = 15$, and likely that an IPHAS star of $i = 18$ would not return a *Gaia* match in this magnitude range, based on the global properties of the two catalogues.

the number of stars within each given narrow magnitude bin that remain. This will also remove some field stars, but under the assumption that the distribution of unrelated stars is positionally uncorrelated the distribution can still be recovered. This may, on the surface, appear to be an asymmetric function; after all, it requires the removal of sources from one catalogue, influenced by the other catalogue while not affecting them in return. However, the key to the symmetry is that f is calculated for both catalogues, and therein lies the symmetrisation of the process.

Determining c is rather more complex. Naively, one might simply record the magnitudes of those stars in catalogue ϕ close enough to the stars in question in catalogue γ to be considered potential counterparts. However, there will be randomly placed unrelated stars that happen to lie close enough to another star to be considered a match, which will then be included in any distributions created. To overcome this interloper problem, a sensible choice would then be to subtract a representative number of stars from each magnitude bin, using f as the distribution to construct the "background". However, as shown in Figure 3.3 for the example of 2MASS sources positionally correlated with *Gaia* sources $15 \leq G \leq 15.25$ at $120 \leq l \leq 125$, $0 \leq b \leq 5$, stars suffer from the crowding out of detections of stars fainter than themselves. The number of faint field stars to be subtracted would therefore be overestimated if the magnitudes of stars close to the chosen objects was used naively.

Instead of considering the closest stars to the sources, the crowding effects can be overcome by considering the brightest sources within a given radial offset, as developed in section 4 of Naylor, Broos, and Feigelson (2013). Using the bright star distribution, which is a density-independent measure, the decrease in the density of fainter objects can be controlled for. Unrelated field objects can then correctly be removed from the distribution, obtaining a more robust counterpart distribution.

However, Naylor, Broos, and Feigelson (2013) only considered a one-sided problem, which effectively put the entirety of the second catalogue into one, very large, magnitude bin. The two-directional case requires the building of $c(m_\phi|m_\gamma)$ for each m_γ to $m_\gamma + \mathrm{d}m$

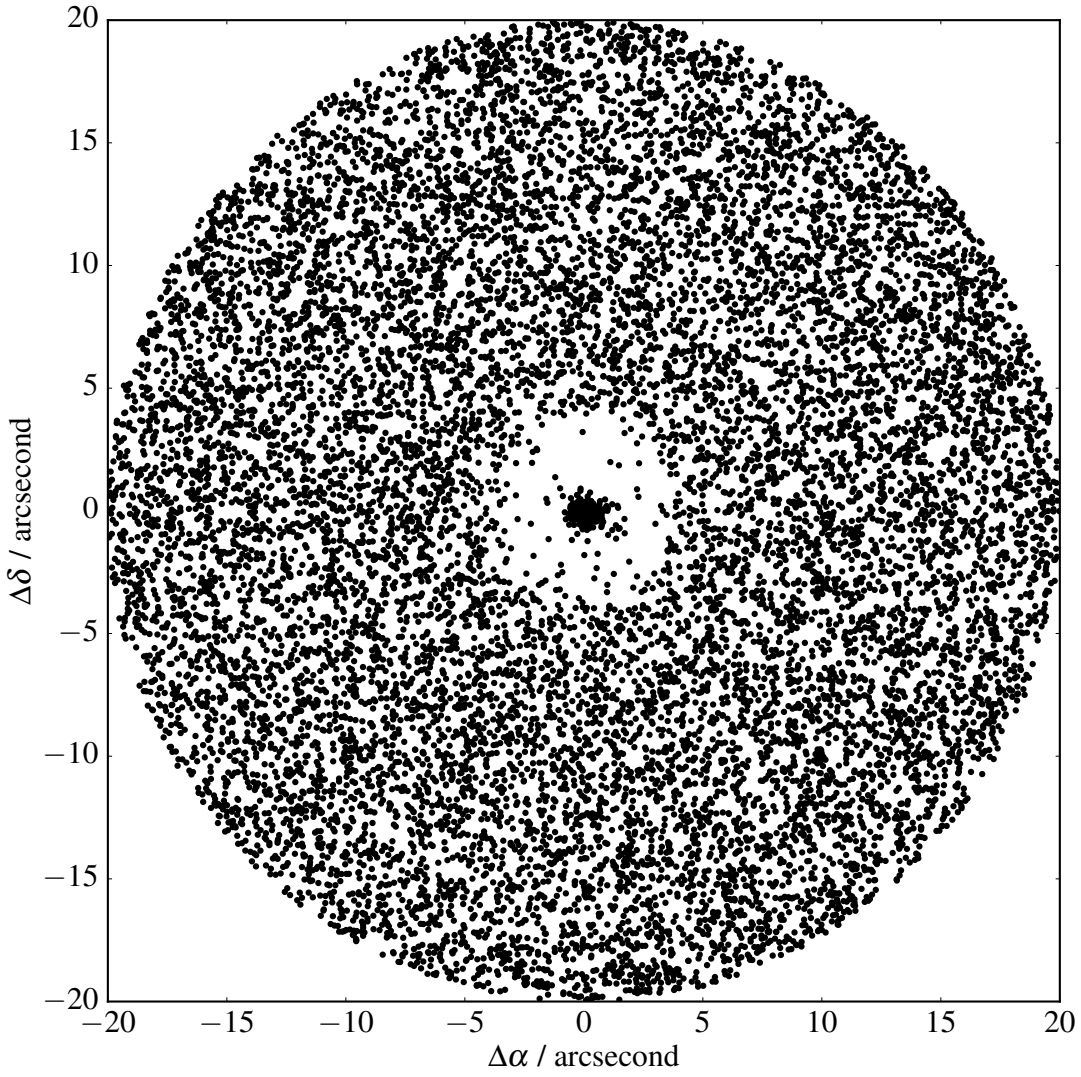


Figure 3.3: The spatial separation of all 2MASS stars within 20 arcseconds of *Gaia* sources $15 \leq G \leq 15.25$, for a $5^\circ \times 5^\circ$ slice of the Galactic plane. Background sources are seen at a constant density surrounding a clump of counterpart stars in the centre. However, the background density decreases within $\lesssim 3.5$ arcseconds due to the crowding out of the fainter background sources by bright counterparts.

bin, in turn. The revised version of equation 16 of Naylor, Broos, and Feigelson (2013) is therefore

$$\begin{aligned} Z_{c\gamma} \cdot c_\gamma(m_\phi|m_\gamma) = Z_\gamma b_\gamma(m_\phi|m_\gamma) \exp\left(A_\gamma N_\phi F_\phi(m_\phi)\right) - \\ \left(1 - Z_{c\gamma} C_\gamma(m_\phi|m_\gamma)\right) A_\gamma N_\phi f_\phi(m_\phi). \end{aligned} \tag{3.40}$$

Here $Z_{c\gamma}$ is the fraction of stars of magnitude m_γ to $m_\gamma + \mathrm{d}m$ with counterparts inside a certain radial distance and Z_γ is the fraction of stars of magnitude m_γ to $m_\gamma + \mathrm{d}m$ with at least one star within the given radius. $b_\gamma(m_\phi|m_\gamma)$ is the distribution of the brightest stars within a radial offset of stars of magnitude m_γ to $m_\gamma + \mathrm{d}m$. A_γ is the average area inside the radial offsets for stars of magnitude m_γ to $m_\gamma + \mathrm{d}m$ and N_ϕ is the number density of unmatched stars in catalogue ϕ. $F_\phi(m_\phi)$ is the integral of the unmatched star distribution for catalogue ϕ, $f_\phi(m_\phi)$, from $-\infty$ to m_ϕ, and $C_\gamma(m_\phi|m_\gamma)$ is the integral of the counterpart star distribution, $c_\gamma(m_\phi|m_\gamma)$, from $-\infty$ to m_ϕ.

There is an equivalent case with the switching of catalogues,

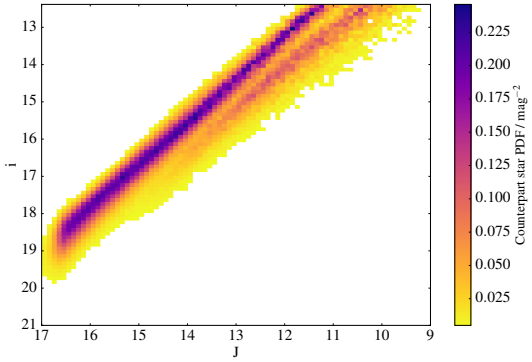
$$\begin{aligned} Z_{c\phi} \cdot c_\phi(m_\gamma|m_\phi) = & Z_\phi b_\phi(m_\gamma|m_\phi) \exp\left(A_\phi N_\gamma F_\gamma(m_\gamma)\right) - \\ & \left(1 - Z_{c\phi} C_\phi(m_\gamma|m_\phi)\right) A_\phi N_\gamma f_\gamma(m_\gamma). \end{aligned} \tag{3.41}$$

These are not truly symmetric (see Figure 3.4 for comparison), because they are, effectively, expressions for $p(a|b)$ and $p(b|a)$; the conditional probabilities of a given b and of b given a, respectively. However, it is easy to obtain the joint probability of a and b by

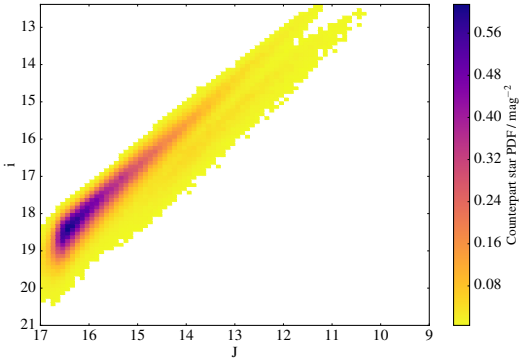
$$p(ab) = p(a|b)p(b) = p(b|a)p(a). \tag{3.42}$$

The symmetrisation of c, from equations 3.40 and 3.41, is therefore

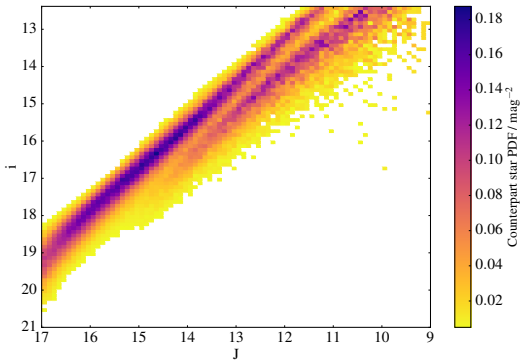
$$c(m_\gamma, m_\phi) = c_\gamma(m_\phi|m_\gamma) \cdot p_\gamma(m_\gamma) = c_\phi(m_\gamma|m_\phi) \cdot p_\phi(m_\phi). \tag{3.43}$$



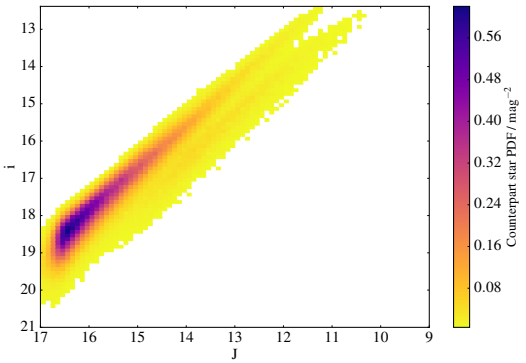
(a) Un-corrected form of c, $c_\phi(m_\gamma|m_\phi)$, using IPHAS as the input catalogue.



(b) Corrected form of c, $c(m_\gamma, m_\phi)$, with IPHAS as the input.



(c) Un-corrected form of c, $c_\gamma(m_\phi|m_\gamma)$, using 2MASS as the input catalogue.



(d) Corrected form of c, $c(m_\gamma, m_\phi)$, using 2MASS as the input catalogue.

Figure 3.4: The effect asymmetry has on the overall counterpart probability density, for the comparison between the J filter in 2MASS and the i filter in IPHAS. Minimum colourmap is 0.005mag^{-2} in all plots. If the symmetrisation step is not taken, the PDF only reflects one catalogue, leading to inconsistent results depending on which catalogue is used as the input. After symmetrisation, however, the PDFs are equivalent. Notation used assumes 2MASS as catalogue γ and IPHAS as catalogue ϕ, following discussion in Section 3.5.

The effects of this additional probability are shown in Figure 3.4, showing that the choice of input catalogue for construction of the magnitude-magnitude relationship does not affect the resulting PDF.

3.6 Application to Photometry

To avoid using bad or unwanted data within individual surveys, I first clean the data using the criteria in Table 1.2. I have chosen three catalogues, *Gaia*, 2MASS, and IPHAS, to highlight two important regimes for probabilistic matching. First, *Gaia* and IPHAS are both optical surveys allowing for ease of comparison, but they have differing dynamical ranges, where IPHAS saturates at a fainter magnitude than *Gaia* but also has a correspondingly fainter completeness limit. Second, the symmetrisation of the matching process means that the handling two catalogues with similar astrometric precision should be possible, which I test with an IPHAS-2MASS cross-match.

While the clean datasets ensure that any spurious artifacts or other non-physical detections in the catalogues are not included, I have also included some flags which remove true stellar detections. This means that the matches do not necessarily include every single source on the sky. Matching two cleaned datasets will result in some unpaired stars which, had poor detections not been removed, should have returned a corresponding detection in the opposing catalogue. This effect is similar to that discussed in Section 3.1, where the saturation of a star in one catalogue and the non-detection of a second star in the opposing catalogue can lead to a nearest neighbour mismatch of the two sources.

One possible solution is to simply remove all stars in all catalogues surrounding a poor quality detection in any catalogue, at the cost of the removal of good quality data. This would allow for a more even matching, where all data were good quality in all potential matches. This, however, unnecessarily removes extra sources from the potential composite catalogue, and thus I chose to only remove the poor quality data. This has the additional advantage for this chapter of leaving these "orphan" stars in the catalogues, which provide a good test of the rejection of star pairings based on their photometry. It

will be seen later in this section that these stars are successfully returned as unmatched field objects.

It is important that any datasets are carefully cleaned for poor quality, spurious or non-required sources. For example, if I had made the choice to keep saturated stars as a "low quality" detection, instead of treating them as being outside the dynamic range of the survey, the effect seen later, of mismatches between nearest neighbour matched sources, the counterpart to which is not included in the opposing catalogue, would be diminished. These saturated objects would still be present in their corresponding dataset and thus chosen over the interloper source as the counterpart to the bright object in the second catalogue.

Furthermore, the removal of specific "classes" of sources (such as non-stellar sources as I have chosen here) will influence the ensemble matches, caused by the effect this would have on the "in situ" photometric PDF creation. The removal of sources of a common physical type – such as galaxies – can avoid confusion in the photometric magnitude-magnitude probability space, aiding the photometric likelihood in distinguishing true and false matches. However, the preferential removal of the majority of a specific class of sources – such as a more distant class of objects, preferentially affected by interstellar extinction – could potentially reduce the number of sources available to populate their parameter space in c and f with sufficient precision, and could affect the match statistics. While this method of deriving photometric likelihoods has an advantage over that proposed by Budavári and Szalay (2008) in its avoidance of relying on theoretical models, it instead relies on sufficient numbers of every class of source in both catalogues to allow for the population of c and f across all magnitudes. Thus it is important to understand any input catalogues to the cross-matching process, their creation pipelines, as well as the scientific aims of the merged dataset, before applying the methodology laid out in Sections 3.3-3.5.

More generally this effect is seen in crowded fields, where one catalogue, with high angular resolution, is matched to another, less able to resolve individual sources. This results in the effect, also discussed later in this section, where the bright resolved

object is matched to the single contaminated source in the opposing catalogue. The faint source in the high resolution catalogue is then returned as an unmatched object. Care must therefore be taken when matching two catalogues of differing resolution to not misinterpret these as stars with corresponding missing detections below the sensitivity of the survey in question. The "completeness limit" of a survey, often quoted as a single magnitude, is therefore highly dependent on the interplay of the resolving power of the survey and the local density of sources.

3.6.1 Integrating Gaussians Under a Circle

Throughout the next two sections I discuss certain "radial" distances, which I define formally here for clarity and notation succinctness. These radial distances, $\mathcal{R}_Y$, are defined as the distance at which a certain fraction (Y) of a circular integral of a two-dimensional Gaussian is enclosed. They are the solution to the equality

$$\iint\limits_{x'^2+y'^2\leq\mathcal{R}_Y^2} G(x', y')\mathrm{d}x'\mathrm{d}y' = Y, \tag{3.44}$$

where G is the convolution of two sources' AUFs (see Section 3.3.2.1 for definition and discussion). The evaluation of the integral of $(f * g)(x, y)$ in a circle defined as $x^2 + y^2 \leq \mathcal{R}^2$ is therefore necessary. To achieve this an identity of the convolution theorem could potentially be used, which states that, for a two dimensional convolution,

$$\iint (f * g)\mathrm{d}x\mathrm{d}y = \iint f\mathrm{d}x\mathrm{d}y \cdot \iint g\mathrm{d}x\mathrm{d}y. \tag{3.45}$$

This step is unnecessary in the specific case of the convolution of two Gaussians, however, as their convolution is itself a new Gaussian. Thus, in order to evaluate the potential convolution integrals it must be possible to express the integral of arbitrarily oriented, elliptical Gaussian distributions inside a circle. First, $\iint f\mathrm{d}x\mathrm{d}y$ shall be expressed as

$$P(\mathcal{R}, \sigma_x', \sigma_y') = \iint\limits_{x^2+y^2\leq\mathcal{R}^2} f_G(x, y, \sigma_x', \sigma_y', \rho)dxdy$$
$$= \iint\limits_{x^2+y^2\leq\mathcal{R}^2} \frac{1}{2\pi\sigma_x'\sigma_y'\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2\sqrt{1-\rho^2}}\left(\frac{x^2}{\sigma_x'^2} + \frac{y^2}{\sigma_y'^2} - \frac{2\rho xy}{\sigma_x'\sigma_y'}\right)\right)\mathrm{d}x\mathrm{d}y. \tag{3.46}$$

As circles are invariant under rotational transformations, it is possible to rotate into the frame of the ellipse, by

$$\begin{aligned}\sigma^2_{\text{major}} \equiv \sigma_x^2 &= \frac{1}{2}\left(\sigma_x'^2 + \sigma_y'^2 + \sqrt{\left(\sigma_x'^2 - \sigma_y'^2\right)^2 + 4\left(\rho\sigma_x'\sigma_y'\right)^2}\right)\\ \sigma^2_{\text{minor}} \equiv \sigma_y^2 &= \frac{1}{2}\left(\sigma_x'^2 + \sigma_y'^2 - \sqrt{\left(\sigma_x'^2 - \sigma_y'^2\right)^2 + 4\left(\rho\sigma_x'\sigma_y'\right)^2}\right)\end{aligned} \tag{3.47}$$

giving a simpler form of

$$P(\mathcal{R}, \sigma_x, \sigma_y) = \iint\limits_{x^2+y^2\leq\mathcal{R}^2} \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right)\mathrm{d}x\mathrm{d}y. \tag{3.48}$$

If

$$a = \mathcal{R}/\sigma_x,\ b = \mathcal{R}/\sigma_y,\ x/\sigma_x = t\cos(\theta),\ y/\sigma_y = t\sin(\theta),$$
$$t'(\theta) = ab/\sqrt{a^2\sin^2(\theta) + b^2\cos^2(\theta)}, \tag{3.49}$$

then this becomes

$$
\begin{aligned}
P(\mathcal{R}, \sigma_x, \sigma_y) &= \int_0^{2\pi} \int_0^{t'(\theta)} \frac{t}{2\pi} \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t \mathrm{d}\theta \\
&= 1 - \frac{2}{\pi} \int_0^{\pi/2} \exp\left(\frac{-a^2 b^2}{2\left(a^2 \sin^2(\theta) + b^2 \cos^2(\theta)\right)}\right) \mathrm{d}\theta.
\end{aligned}
\tag{3.50}
$$

At this point the integral could be considered in terms of the circular coverage function and Bessel function theory (Gray, Mathews, and Macrobert, 1895), and, defining the coverage function as

$$
Q(x, z) \equiv \exp\left(-z^2/2\right) \int_0^x \exp\left(-t^2/2\right) \mathrm{I}_0(zt)\, t\, \mathrm{d}t, \tag{3.51}
$$

where I_0 is the modified Bessel function of the first kind with order zero, the integral of the ellipsodial Gaussian distribution inside a circle can be expressed as

$$
P(\mathcal{R}, \sigma_x, \sigma_y) = Q\left(\frac{\mathcal{R}}{2}\left(\frac{1}{\sigma_x} + \frac{1}{\sigma_y}\right), \frac{\mathcal{R}}{2}\left|\frac{1}{\sigma_x} - \frac{1}{\sigma_y}\right|\right) - Q\left(\frac{\mathcal{R}}{2}\left|\frac{1}{\sigma_x} - \frac{1}{\sigma_y}\right|, \frac{\mathcal{R}}{2}\left(\frac{1}{\sigma_x} + \frac{1}{\sigma_y}\right)\right). \tag{3.52}
$$

However, it is acceptable to leave the probability in the form given in equation 3.50, as it is more compact and simpler to evaluate in typical use.

3.6.2 Reducing Computational Complexity

Equation 3.35 is too computationally expensive to treat the entirety of a catalogue as one set, as discussed in Section 3.3.3.2. I reduce the complexity by initially assuming that there is no overlap between stars drawn from the same catalogue, which I shall refer to as "internal independence". However, the chance of a star from catalogue ϕ being positionally close to two stars from catalogue γ must be accounted for, even if those original stars are not positionally overlapping one another. Such "external dependencies" would not allow for the treatment of stars in catalogue γ as independent and force them to be considered

as part of a larger set. This assumption is borne out in the one-directional case considered by Naylor, Broos, and Feigelson (2013), in which they were able to assume their X-ray dataset was internally independent, but, due to the multiplicity of the potential matches, the IR data were not independent of one another. Here I am simply generalising this to both catalogues, creating "groups" of both sets of, e.g., X-ray and IR, detections. I have therefore relaxed the assumption that internal independency holds for one of the catalogues, but must break the matches up into groupings which have inter-group independency, for computational purposes.

To break the matches into independent groupings requires first iterating over the entirety of one catalogue, assigning as potential counterparts to each star those stars in the other catalogue which appear within a certain "merging radius". These potential counterpart lists are merged in cases, as previously, where two stars could potentially match to the same star in the opposing catalogue. These mergers give a complete list of "islands" which are independent of each other but must be considered jointly within. I am extremely conservative with my rejecting of potential counterparts, using a large merging radius.

To calculate the radius at which objects must be considered close enough to be related, the star at the 95^{th} percentile uncertainty ellipse area – πab – is found for each catalogue. This gives uncertainties that avoid significant outliers, but that are larger than those of the vast majority of the survey. The semi-major and semi-minor axes of those stars are then used to construct G. Stars are defined to be positionally close to one another if they are separated by less than $\mathcal{R}_{0.997}$ ($\simeq 3.4\sigma$ for a circular, two-dimensional Gaussian), the critical merging radius.

Each island is then fed into equations 3.34 and 3.35, and the most probable arrangement is accepted, with stars being assigned as counterparts or unmatched stars. This permutation can then either be accepted or it can be rejected as uncertain depending on whether its probability lies above a certain threshold. For example, the most likely permutation can be accepted, no matter the probability; permutations can be accepted with

$P > 0.5$, where the highest probability permutation outweighs all other permutations; or a more strict criterion can be used, requiring $P > 0.8$ (e.g., Broos et al., 2013). The probabilities in this section are accepted where the overall permutation probability $P > 0.5$; i.e., where the most likely permutation is more likely than all other options combined.

3.6.3 Constructing f and c computationally

To calculate f, a large section must be "cut out" around each catalogue γ star in catalogue ϕ, to avoid any possibility of introducing the true counterpart to the unmatched probabilities. However, due to the large variations in precision for detections, each star must be considered individually when avoiding potential counterparts. When masking a given star in catalogue γ, any stars in catalogue ϕ within a certain distance are ignored. This distance is found by finding the star in catalogue ϕ in the same "island" as the catalogue γ star in question with the largest astrometric uncertainties. The two stars' AUFs are then used to create a new G distribution, and find $\mathcal{R}_{0.9}$. It is this radius inside which catalogue ϕ objects close to the catalogue γ star are ignored. $Y = 0.9$ was chosen as a tradeoff between two requirements. First, the contamination from counterparts appearing in the uncorrelated sample should be at a minimum. It is nominally at the 10% level but mitigated by the fact that G always uses the largest possible uncertainties. Second, if possible low number statistics should be mitigated against, avoiding overly large "cut out" radii caused by the integration of G to large distances. In addition to calculating f_γ and f_ϕ, N_γ and N_ϕ are calculated from the area the catalogue covers after the star masks were applied, subtracting the total area masked by the calculated radial offsets.

To construct c, equation 3.40 is used, and therefore the building of distributions of b, the bright star distribution, is required. For this, I define radii for each star in a given catalogue in a similar way to when f was constructed, except I use $\mathcal{R}_{0.63}$, the 0.6×FWHM optimal result from Naylor (1998). This radius trades off between minimising the effects of unmatched stars in the distributions while ensuring there are still enough counterparts to ensure good number statistics. N_c was calculated by integrating each $Z_{c\phi} \cdot c_\phi(m_\gamma|m_\phi)$ to

obtain $Z_{c\phi}$, because each c_ϕ slice should be normalised if the b_ϕ slice and f_γ are normalised. This then gives us, for the magnitude slice, the fraction of stars with counterparts within $\mathcal{R}_{0.63}$. To obtain the overall fraction of stars with counterparts, this fraction must be divided by the fraction expected, $Y = 0.63$. Once the fraction of input objects which have counterparts is found, the number density of counterparts can be obtained by multiplying by the number density of sources in the small magnitude slice. Repeating this for all magnitudes, the density of counterparts for each input magnitude slice is summed to obtain the total counterpart number density, $N_{\rm c}$.

Throughout this section I will be comparing number densities of matches, for both the matched counterparts and unrelated field stars. For the one dimensional density these are simply the number of objects with a magnitude m_γ to $m_\gamma + \Delta m_\gamma$, T, divided by bin width Δm_γ. In the two dimensional case the number density is the number of objects with magnitude m_γ to $m_\gamma + \Delta m_\gamma$ and magnitude m_ϕ to $m_\phi + \Delta m_\phi$, T, divided by bin widths $\Delta m_\gamma \Delta m_\phi$. I will consider three sources of counts: the probability-based counterpart matches ($T_{\rm prob}$), the nearest neighbour-based matches ($T_{\rm prox}$; "proximity" matches), and the probability-based unmatched objects. These number densities, while not normalised, are comparable to the PDFs c and f. The number density of counterparts is related to $AN_c c$, where A is the area of sky under consideration, while $AN_\phi f_\phi$ is the equivalent field star number density.

3.6.4 Probabilistic Matches

Having constructed both the astrometric uncertainty functions and the counterpart and unmatched star magnitude PDFs, the matching between catalogues can now be done. For the test cases, the two catalogues were extracted for a 25 square-degree area of the sky, $120 \leq l \leq 125$, $0 \leq b \leq 5$, and any stars which did not contain at least one filter flagged as a detection (either good or low quality) were discarded. Then c and f were constructed for each filter – i for IPHAS, J for 2MASS, and G for *Gaia* – along with the corresponding number densities.

3.6.4.1 IPHAS vs *Gaia*

I begin with the case of two optical catalogues, *Gaia* and IPHAS. *Gaia* saturates at a brighter magnitude than IPHAS, while IPHAS has a fainter completeness limit, which allows me to test my matching in the case of differing dynamical ranges. Figure 3.5 shows the distributions of counterpart and unmatched stars for *Gaia* G and IPHAS i, comparing a 3 arcsecond nearest neighbour match to the probabilistic matching, accepting only those islands in which the most likely permutation is more probable than all other permutations. This nearest neighbour match is larger than the maximum island acceptance radius, resulting in a small number ($\lesssim 1\%$) of cases where there is a nearest neighbour match but no probability-based match based on the rejection of association during the island creation. However, these objects are rejected on both astrometric and photometric grounds, and I do not consider them further.

Several things need to be checked, using Figure 3.5, before the method can be assumed to correctly pair the correct objects. First, stars in *Gaia* that correspond to the saturated region in IPHAS should be returned as unmatched stars. The matched stars returned are shown as solid black lines in the side panels of Figure 3.5, and a clear rejection of any match for stars of $G \lesssim 13$ (i.e., those detections saturated in IPHAS) can be seen. Second, given the nature of matching two catalogues in the optical, all stars should be returned as being matches in the dynamical range of the two catalogues. Comparing the matches in $13 \lesssim i \lesssim 19$, the matches can be contrasted with a naive 3 arcsecond nearest neighbour match, shown as the solid black lines and red dot-dashed lines in the side panels of Figure 3.5 respectively. The probability-based matches return almost all of the nearest neighbour-based matches, as expected. Those unmatched objects in this region of overlapping dynamical ranges between the two catalogues are unexpected, with approximately one in five objects in either catalogue in this brightness range failing to return a match. However, over 80% of these objects have no counterpart in the opposing catalogue within 5 arcseconds (Section 3.6), and are simply objects whose counterpart was rejected from the cleaned catalogues by my selection criteria (Table 1.2). The remaining

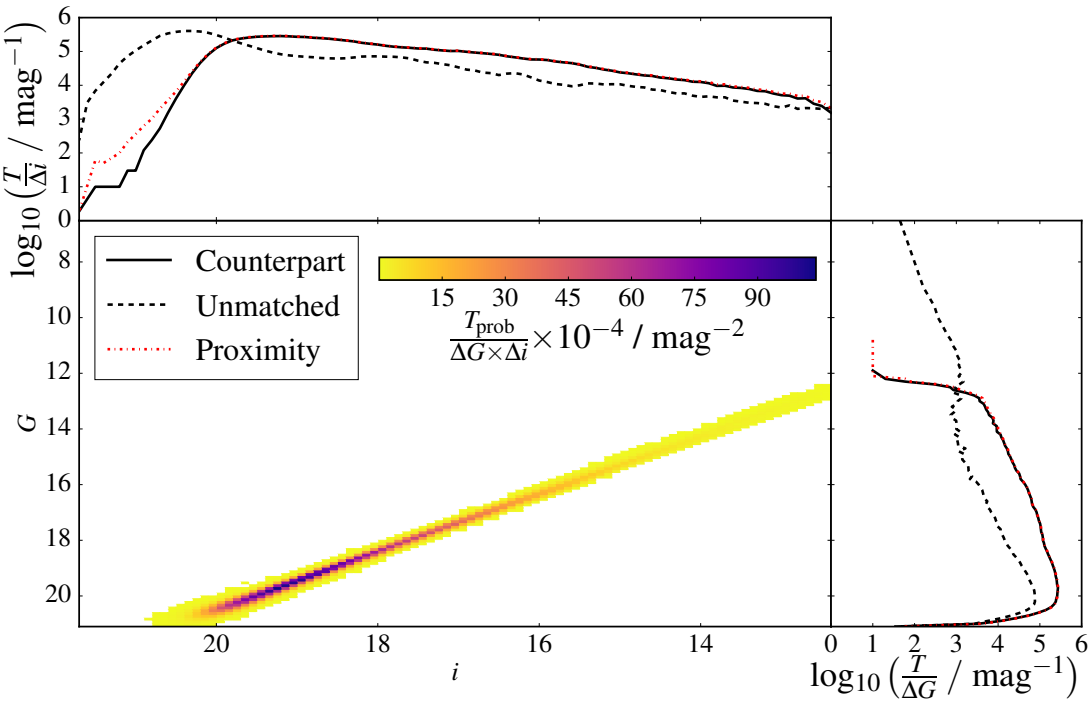


Figure 3.5: The distributions for the probability matching of *Gaia* and IPHAS in a 25 square degree region of the Galactic plane, in the G and i filters respectively. The middle panel shows a 2D histogram of probability-based counterparts in each small magnitude-magnitude bin. As expected from two similar optical passbands, the counterpart magnitude trend is roughly linear with decreasing brightness. The top and side panels show the number density of sources in each filter individually (i.e., the total number of stars returned as counterparts with a specific G magnitude) in the solid black lines. Also shown in the inset figures are the unmatched star number densities (dotted black lines) and a 3 arcsecond nearest neighbour-based match ("proximity" matches; red dot-dashed lines). The counterparts returned by nearest neighbour- and probability-based matches agree for most magnitude ranges. However, in the case of nearest neighbour matches an increase in the number of bright *Gaia* counterparts that match to faint IPHAS objects is seen, which the probability-based match rejects. Colourmap only displayed for those bins with densities $\geq 500\text{mag}^{-2}$.

20%, which do have a nearest neighbour match, are discussed later. Third, any potential mismatches between faint IPHAS objects and brighter *Gaia* stars should be removed. Fainter than $i = 20$, a decrease in the number of counterparts returned by the probabilistic match can be seen, compared to the traditional nearest neighbour match (black solid lines vs red dashed lines in inset figures to Figure 3.5). One in four nearest neighbour matches is rejected as a probabilistic match fainter than $i \simeq 20$, a minority of which are systematically perturbed true matches and also discussed below. The loss rate increases by $i \simeq 21$ to four in every five nearest neighbour match pairs being assigned as unrelated, unmatched objects by the probability-based match. These rejections are mostly IPHAS objects too faint in G to be detected, but serendipitously close to an unrelated bright *Gaia* object, flagged in IPHAS. They have therefore been picked up as an unphysical match, and would be paired without the addition of the magnitude information.

A small fraction of objects are returned as field stars at brighter magnitudes that nearest neighbour matching assigns as counterparts. This population should be considered in more detail. Figure 3.6 shows the difference in the number density of probability- and 3 arcsecond nearest neighbour-based matches. In the bright dynamic range of *Gaia*, $12 \leq G \leq 17$, the typical loss of objects is $\simeq 3\%$. However, this loss rate is across all IPHAS magnitudes, and includes $\lesssim 1\%$ loss rate (i.e., one third of the total number of lost matches) of objects in the high counterpart density region of the magnitude-magnitude diagram. The rejections where the IPHAS magnitudes do not agree with the *Gaia* brightness are reasonable and show the additional magnitude information correctly rejecting unlikely counterparts. However, the 1% of rejections where the i and G magnitudes lie in the narrow range of accepted counterparts in both filters ought to be paired, and require further consideration.

When considering these unexpected rejections I can highlight the effect the magnitude information has on the counterpart matching scheme. However, before I am able to do so I must re-introduce the likelihood ratio (Sutherland and Saunders, 1992; Fleuren et al., 2012; Brusa et al., 2005; etc.), but split it into the photometric and astrometric

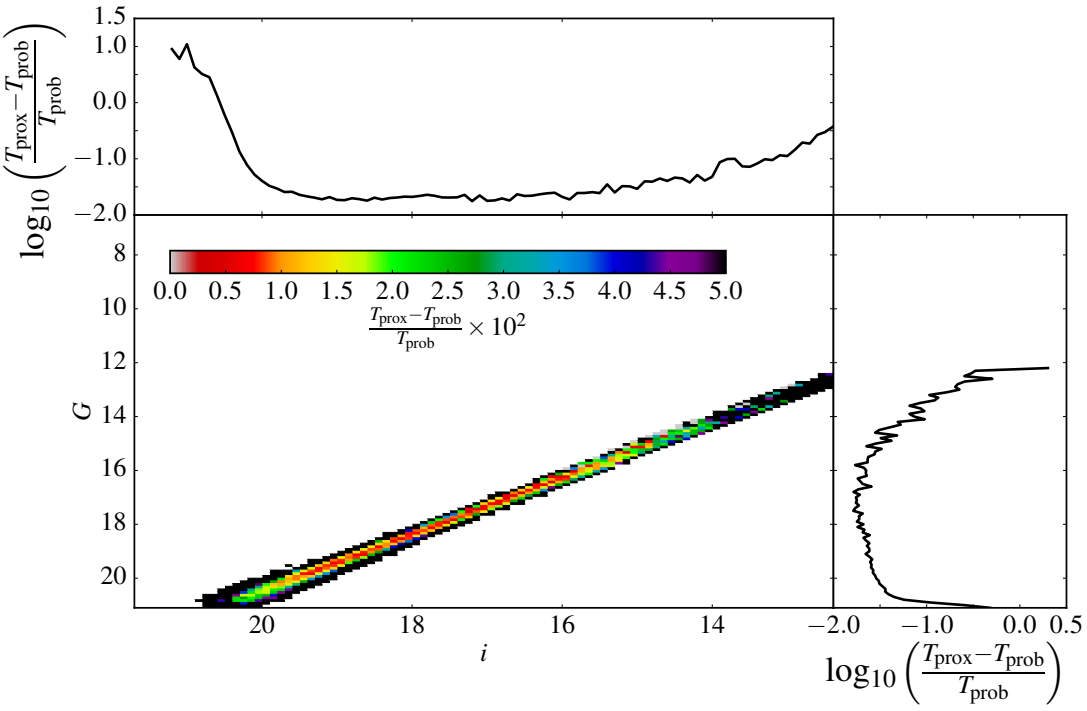


Figure 3.6: The relative difference in number of objects returned for an IPHAS-*Gaia* cross-match for 25 square degrees of the Galactic plane. Main panel shows the relative difference between the probability- and nearest neighbour-based ("proximity") matches for each small magnitude-magnitude bin, while the inset panels show the relative difference for each magnitude. At bright magnitudes a consistent rejection of matches occurs for $\lesssim 3\%$ of objects. However, at fainter magnitudes ($i \gtrsim 20$) rejection of nearest neighbour matches occurs at a higher rate, caused in part by the assumption that the IPHAS AUF is purely Gaussian. The assumption of Gaussianity will cause the rejection of those objects in the non-Gaussian tails caused by systematic perturbations such as contamination due to faint, unresolved objects in the IPHAS PSF (see Chapter 2). Bins shown in main panel are the same as those which met the criterion in Figure 3.5.

components of, e.g., equation 3.32. The photometric likelihood ratio, η, logarithmically balances the likelihood of matching magnitudes against the likelihood of the two stars being photometrically unmatched, given by

$$\eta \equiv \log_{10}\left(\frac{c(m_\gamma, m_\phi)}{f_\gamma(m_\gamma) f_\phi(m_\phi)}\right). \tag{3.53}$$

Equivalently, the astrometric likelihood ratio, ξ, is the logarithm of the comparison between the astrometric counterpart likelihood and the likelihood of the two objects being unrelated astrometrically, defined as

$$\xi \equiv \log_{10}\left(\frac{N_c G}{N_\gamma N_\phi}\right). \tag{3.54}$$

Consider Figure 3.7, which shows the main locus of those objects matched successfully by the probabilistic matching process (red solid contours). Also shown, in black dashed contours, is the area occupied in the ratio-ratio space by those objects that are returned by a nearest neighbour-based matching process but not by a probability-based match (i.e., those objects in Figure 3.6). The vast majority of objects lost between the two processes are not lost due to low photometric chance. In fact, the contours lie in roughly the same region in η, but the lost objects have likelihood ratios six orders of magnitude lower in astrometry, compared to the main matched set. In both cases, the average improvement to the likelihood ratio that η gives is approximately a 10-fold increase in probability. These high photometric likelihood but low astrometric likelihood objects are those whose astrometric positions are perturbed by systematic effects. They are perturbed to such a degree that they fall outside the maximum separation allowed by a Gaussian AUF (see Chapter 2). They are still within 3 arcseconds, however, and are therefore still picked up by a nearest neighbour match. This lowers their astrometric likelihood ratio until they become more likely unrelated objects than counterparts to the same source, as defined by the dotted line $\xi + \eta = 0$. These "incorrect" losses can be distinguished from truly rejected nearest neighbour matches by comparing both the photometric and astrometric likelihood ratios. While those matches that should not have been lost are only lost on astrometric

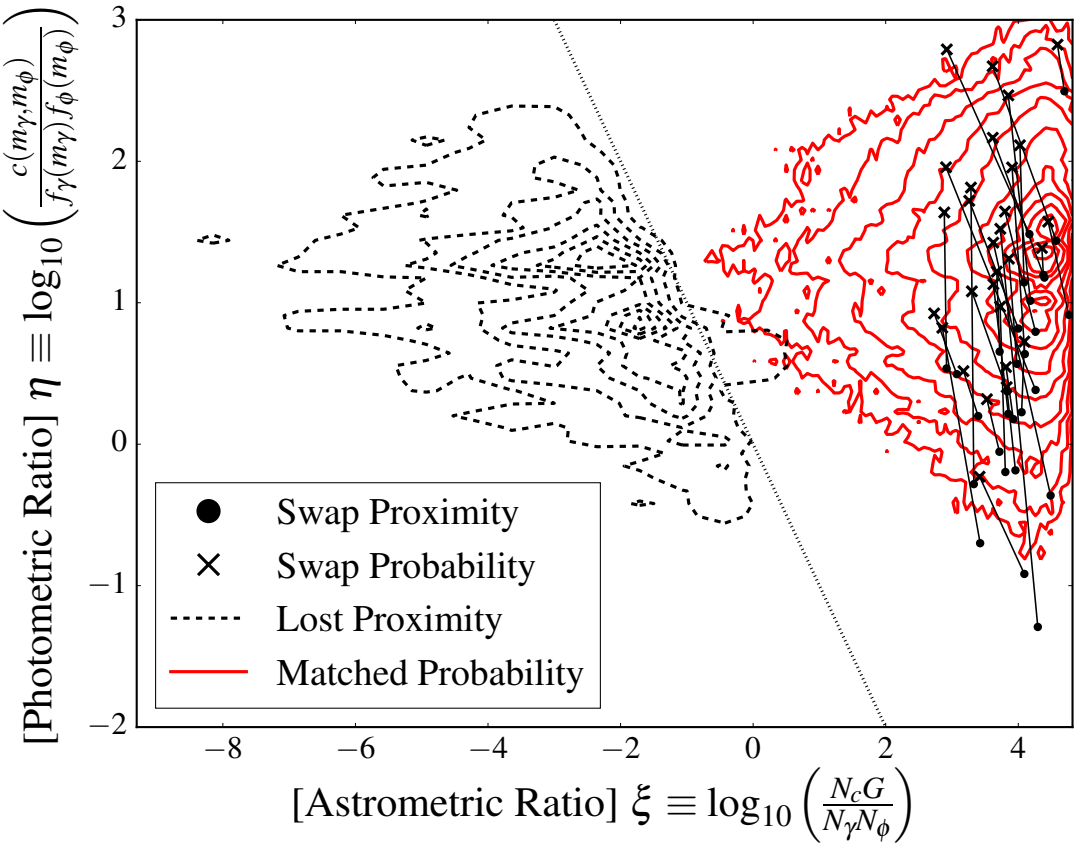


Figure 3.7: The relative likelihoods of matched IPHAS and *Gaia* stars, for a 25 square degree section of the Galactic plane. Here the two likelihood ratios, photometric and astrometric, for the matches between the datasets are being compared. Red solid contours show the area of the plot occupied by the majority of the probability-based matches, while the black dashed contours show the area occupied by objects which were nearest neighbour matched to 3 arcseconds, but failed to return a probability-based match. Additionally, the connected lines are the cases where stars were nearest neighbour matched to one object, but returned a different probability-based match. These likelihood ratios are denoted by crosses for the probability-based match, circles for the nearest neighbour-based match, and are connected by a solid black line. Dotted line $\eta + \xi = 0$ represents a combined likelihood ratio of unity; equal chance between the two hypotheses.

grounds, a serendipitous nearest neighbour match has both poor photometric and astrometric likelihood ratios. A few objects are also seen whose astrometric likelihood ratios are very high, but have photometric ratios slightly below one. These are the rare cases where objects coincidentally have magnitudes more typical of unrelated field objects (e.g., uncommon stellar types, non-stellar sources which have not been removed from during the data reduction process, etc.). However, their sky proximity is so overwhelmingly unlikely if they were unrelated that they simply must be detections of the same original object.

The few cases in the set where one star has "skipped" over its closest neighbour and been matched with a nearby, but more distant, counterpart can also be considered, similar to the example laid out in Section 3.2. In these cases the sky separation has increased, decreasing slightly the probability density G, but trading off against a large increase in photometric likelihood, as seen in Figure 3.7 as the connected lines. This demonstrates the value of the additional information gained by using the photometry, allowing for the avoiding the pairing of two unrelated but serendipitously located objects.

3.6.4.2 IPHAS vs 2MASS

Next, the matches between IPHAS and 2MASS can be compared. For this matching process, however, there is not a one-sided astrometric precision between the catalogues, because both IPHAS and 2MASS both have similar, $\simeq$0.05 arcsecond positional precision in their bright, non-saturated regimes. This means that neither catalogue would be the obvious choice to map the other onto in an asymmetric matching fashion. It is therefore an important test of the symmetrisation of the photometric probabilities to the two-directional case.

Both counterparts and unmatched stars are successfully recovered by the probability-based matching scheme, shown in the side panels of Figure 3.8 as solid and dashed black lines respectively, in the correct magnitude ranges as with the IPHAS-*Gaia* case above. Here a larger spread is seen in accepted magnitudes in one catalogue for a given brightness in the other (i.e., a larger spread in $i - J$ colour), shown in the main panel of Figure 3.8 as

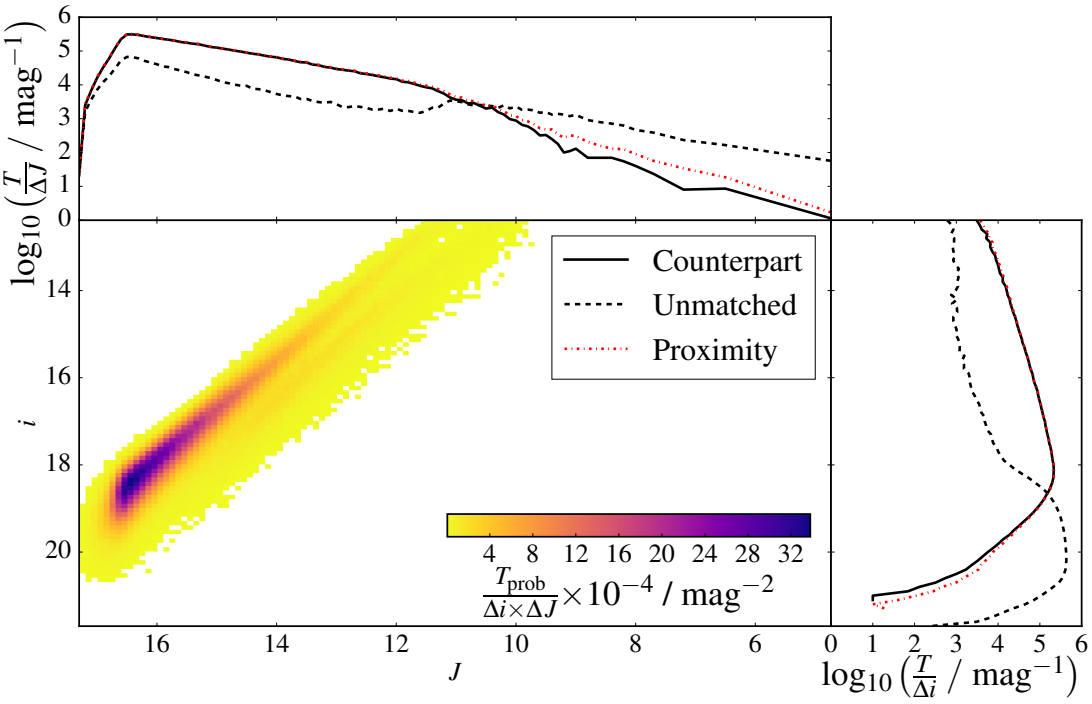


Figure 3.8: The distributions of probability matched counterpart stars for 2MASS and IPHAS in a 25 square degree region of the Galactic plane, in the J and i filters respectively. Figure layout and colourbar are the same as Figure 3.5. Note the comparison to the nearest neighbour-based matches, where stars $J \leq 10$ are incorrectly assigned as matched to stars $i \geq 20$.

an increase in the number density of matches across a larger area of magnitude-magnitude parameter space. This is due mostly to the effects of differential extinction affecting the optical and near infra-red detections to differing degrees. Both IPHAS and now 2MASS contribute to the non-Gaussian tails in the wings of the AUFs. This means that the matches still suffer from the rejection of several percent of likely counterparts at $i \simeq 18$, in a similar effect to that described in Section 3.6.4.1. Additionally, an increase in the rejection of the pairing of faint IPHAS objects with bright 2MASS objects is seen, as shown in the larger differences between the probability- (solid black) and nearest neighbour-based (red dot-dashed lines) counterpart distributions in the side panels of Figure 3.8.

An effect is seen which is not seen in the *Gaia*-IPHAS case. In the case of the likelihood ratio comparison, there are some cases where both the astrometric and photometric likelihood ratios are increased by changing to a more distant counterpart, compared with that returned from nearest neighbour matching. These are the cases where a very faint object, which therefore has large astrometric uncertainties, is slightly closer to a bright object than another bright, and therefore astrometrically precise, object. This decrease in astrometric uncertainty leads to an increase in G, and thus ξ. The previously seen increase in η is still observed, as the brighter object is correctly assigned as the counterpart.

3.6.4.3 The Likelihood Ratio As a Transient Detector

When discussing the match rejections of IPHAS and *Gaia* in Section 3.6.4.1 I distinguished between incorrect losses, caused by assumptions about the description of the AUF, and true match rejections, caused by random chance alignment of two uncorrelated sources. The astrometric-photometric likelihood ratio space (Figure 3.7) can be used for more than this differentiation, however. True matches have high astrometric and photometric LRs, false losses have low astrometric LRs but high photometric LRs, and true rejections have high astrometric LRs but low photometric LRs. The fourth region of the LR-LR plot can be considered: high astrometric LR and low photometric LR. These objects lie very close

in astrometric separation, yet their photometric colour is vastly different to that of the statistical bulk of the sources surrounding them.

Based on their magnitudes in one catalogue, these sources would not be expected to have their detected brightness in the second catalogue, and yet their positions are overwhelmingly unlikely to be so close by random chance. This region of the likelihood ratio space can therefore be used to probe for transient objects. An example of this is an object recorded in the APASS DR9 catalogue at $V = 11.4$. This detection suggests a reasonably bright source, likely a relatively nearby star. However, the object was also observed during the IPHAS campaign. A mere 0.08 arcseconds from the APASS position is a source with an r-band detection of 21.2, a full 10.6 magnitudes fainter than the detection in the APASS catalogue. However, its separation, less than a tenth of an arcsecond, gives such a high astrometric likelihood ratio that these observations simply must be of the same astrophysical object.

The object in question is Nova Cep 2013. Discovered in February 2013 at a V-band magnitude of 11 (Munari et al., 2013), it was serendipitously observed during its decline as part of the AAVSO Photometric All Sky Survey (APASS; Henden and Munari, 2014) observation campaign. The IPHAS DR2 campaign ran between 2003 and 2012, and thus observed the source before its Nova outburst. This object therefore has an extremely high astrometric likelihood ratio, and yet photometry that is attempting to reconcile a V magnitude of 11.4 with an r magnitude of 21.2, resulting in an impossibly low photometric likelihood ratio. This parameter space is therefore an ideal indicator for such transient outbursts. The matching of two catalogues with sufficient temporal resolution – or the same observation campaign in two different epochs – could resolve these brief changes in source fluxes by leveraging this imbalance in the astrometric and photometric information.

3.6.5 Summary

In this section I applied the probability-based matching scheme to three test photometric catalogues, for the cases of *Gaia* matched with IPHAS and IPHAS matched with 2MASS.

I used the method as described in Sections 3.3 through 3.6.3. In both cases, I confirm the method correctly returns the majority of nearest neighbour-based matches.

I discussed the key areas of the magnitude-magnitude space where the number of probabilistic matches deviates from the number of nearest neighbour matches. I concluded that the method is correctly rejecting some faint, nearest neighbour matched objects and assigning a brighter, more distant object as the counterpart. Additionally, I rejected some nearest neighbour matches which are the proximity pairing of two different objects, matched accidentally. One object is lost (through, e.g., saturation or a poor detection) in catalogue γ but within the dynamical range of catalogue ϕ, while the other object is too faint to be included in catalogue ϕ but detected with good signal in catalogue γ. While I also rejected some likely counterparts (i.e., two detections with similar magnitudes in similar passbands which would be expected to be the same source), I showed these failed matches are lost based on their astrometry rather than their photometry. The assumption of pure Gaussian AUFs leads to unphysically small astrometric probabilities when objects are systematically perturbed to large separations relative to their astrometric uncertainties. The factor of approximately 10 increase in probability introduced with the addition of the photometric likelihoods is simply unable to overcome such low astrometric likelihood ratios.

In all cases, the additional parameter space from the magnitude information contributes to the resultant posterior probabilities. However, if the choice is made to model the probability density of star separations in detail, rather than using a simple cut-off radius, then it is critical that the AUFs are modelled properly. Correct AUF descriptions would minimise the rate of false non-pairings, allowing the photometric probabilities to distinguish between true and false matches.

3.7 Extension to Multiple Catalogues

So far, in Sections 3.2 to 3.6, I have only considered the case where one catalogue is being matched against another. However, oftentimes multiple catalogues should be matched

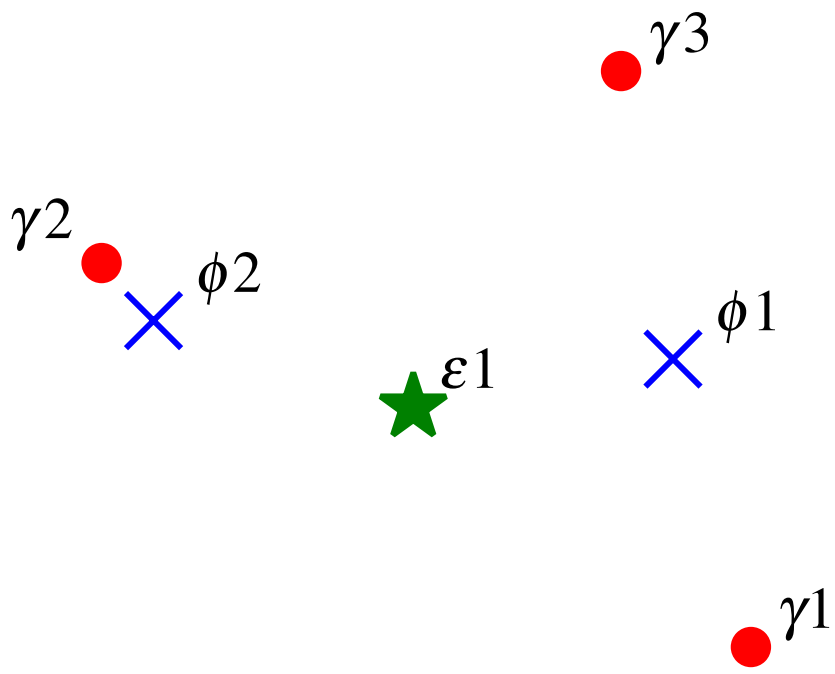


Figure 3.9: Figure showing an arrangement of potential matches from three theoretical catalogues. In this scenario one star is seen in all three catalogues as $\gamma 1$, $\phi 1$, and $\epsilon 1$ respectively; $\gamma 2$ and $\phi 2$ are the same star recorded in two catalogues; and a third star, $\gamma 3$, is only seen in one catalogue. Catalogue γ sources are denoted by red circles, catalogue ϕ sources are shown as blue crosses, and the single green star source is from catalogue ϵ.

to each other, to extend the wavelength coverage. Imagine a hypothetical scenario for a three-catalogue match. Shown in the schematic in Figure 3.9 are three example stars, observed in three example catalogues. Catalogue γ observed three stars in the small field of view in consideration, denoted $\gamma 1$, $\gamma 2$, and $\gamma 3$, shown as red circles. Catalogue ϕ, shown as blue crosses, observed two of the stars: $\phi 1$ and $\phi 2$. Finally, the third catalogue ϵ only recorded a measurement for $\epsilon 1$, shown in Figure 3.9 as the green star.

It is potentially feasible to iterate all possible permutations of this set, asking what the probability is that, e.g., stars $\gamma 2$ and $\phi 2$ are counterparts to each other, star $\gamma 3$ is uncorrelated and stars $\gamma 1$, $\phi 1$, and $\epsilon 1$ are all counterparts of the same object. Considering all possibilities would require extensions to c, asking what the likelihood of counterparts having magnitudes $m_{\gamma 1}$, $m_{\phi 1}$, and $m_{\epsilon 1}$ was, as well as an extension to G, given now as

$$G'(\Delta x_{\gamma 1\phi 1}, \Delta y_{\gamma 1\phi 1}, \Delta x_{\gamma 1\epsilon 1}, \Delta y_{\gamma 1\epsilon 1}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h_\gamma(x_{\gamma 1} - x_0, y_{\gamma 1} - y_0) h_\phi(x_{\phi 1} - x_0, y_{\phi 1} - y_0) \times h_\epsilon(x_{\epsilon 1} - x_0, y_{\epsilon 1} - y_0)\, \mathrm{d}x_0\, \mathrm{d}y_0, \tag{3.55}$$

where h_γ, h_ϕ and h_ϵ are the astrometric distributions of the three catalogue respectively.

However, the complexity of the problem increases geometrically, and it quickly becomes impractical to treat even three catalogues simultaneously. In cases where more than two catalogues are required, sequential matching, starting from the two most astrometrically precise catalogues and working towards the least precise astrometry, is recommended. Starting with a match between catalogues γ and ϕ, catalogue $\gamma\phi$ is created, which contains matches between both catalogues, unmatched catalogue γ objects and unmatched catalogue ϕ objects. Subsequently catalogue $\gamma\phi$ is then taken and matched with catalogue ϵ, creating a catalogue which contains matches between γ, ϕ, and ϵ; γ and ϵ matches; ϕ and ϵ matches; γ and ϕ matches; and objects in catalogues γ, ϕ, and ϵ which do not match to either of the other two catalogues.

For example, a composite catalogue might be required with optical detections (e.g., IPHAS), near-IR sources (e.g., 2MASS), and detections at longer wavelengths (e.g., *Spitzer*; Werner et al., 2004). In this instance the first match might be IPHAS and 2MASS (see Section 3.6.4.2), creating the first sequential composite cross-match catalogue, with *Spitzer* then matched with this new catalogue. However, when matching the second time, any hypothesis where any sources paired during the IPHAS-2MASS match are not paired have been removed from the normalisations (e.g., equation 3.9). However, the choice can be made to only accept high probability classifications from previous iterations of the sequential matching (see Section 3.6.2 for more details). These relatively certain classifications will have low probabilities of any other hypothesis, and the exclusion of the hypothesis of previous IPHAS-2MASS matches being unrelated will have little impact

on the conclusions drawn. Thus the complexity of a multi-catalogue cross-match can be reduced into several two catalogue cross-matches.

The only concession that has to be made is in the careful treatment of equation 3.55 (cf. equation 3.24 for the original two-catalogue case). Equation 3.55 is not easily split into sequential terms, and in order to do so it is necessary to "update" the position of a counterpart pair merge after each cross-match, which is why it is recommended that the most precise catalogues are used initially. The weighted mean position of the two matched stars can then be used as the new position. Updating the position of the source in this way is comparable to section 5.1 of Pineau et al. (2017), although since one cannot guarantee Gaussianity of the distributions (see Chapter 2) this becomes

$$x_{\text{new}} = \frac{\iint_{-\infty}^{+\infty} h_\gamma(x_{\gamma 1}-x_0, y_{\gamma 1}-y_0) h_\phi(x_{\phi 1}-x_0, y_{\phi 1}-y_0) x_0 \,\mathrm{d}x_0 \,\mathrm{d}y_0}{(h_\gamma * h_\phi)(x_{\gamma 2} - x_{\phi 2}, y_{\gamma 2} - y_{\phi 2})} \tag{3.56}$$

with analogous arguments for y_{new}. While it is relatively easy to update the position of the star in the new cross-matched catalogue, it is less straightforward to handle the updated AUF. I therefore recommend simply using the appropriate covariance matrix and AUF of the most positionally precise of the two merged stars.

In the era of increasingly precise datasets, such as *Gaia*, the complication of sequential matching becomes increasingly negligible, as equation 3.55 simply returns

$$\begin{aligned} &G'(\Delta x_{\gamma 1 \phi 1}, \Delta y_{\gamma 1 \phi 1}, \Delta x_{\gamma 1 \epsilon 1}, \Delta y_{\gamma 1 \epsilon 1}) \\ &\quad = h_\phi(x_{\phi 1} - x_{\gamma 1}, y_{\phi 1} - y_{\gamma 1}) h_\epsilon(x_{\epsilon 1} - x_{\gamma 1}, y_{\epsilon 1} - y_{\gamma 1}) \end{aligned} \tag{3.57}$$

in the limit of $h_\gamma(x_{\gamma 1} - x_0) \to \delta(x_{\gamma 1} - x_0)$. Effectively, this simply asks the probability of the two other catalogues being drawn from the order-of-magnitude more precise third position.

3.8 Reduction to One-Sided Case

I have presented a symmetric approach to the probability-based matching procedure treated asymmetrically by several previous authors (e.g., Sutherland and Saunders, 1992; Naylor, Broos, and Feigelson, 2013; Rutledge et al., 2000). To verify the validity of the formalism, I must check that the equations reduce to the one-sided set of equations in the correct limits. As my formalism is based upon that of Naylor, Broos, and Feigelson (2013), I shall confirm that I can recover their equations in this section.

The differences introduced in equations 3.34 and 3.35, compared with equations 6 and 7 of Naylor, Broos, and Feigelson (2013), come from the reduced dimensionality of the problem, as well as several underlying assumptions. If it were possible to treat, e.g., X-ray sources as independent entities, each with a unique set of potential counterparts, the larger catalogue could be broken up into smaller ones, each of which only containing one source, resulting in an effective catalogue length of one. This is equivalent to assuming the catalogue has internal independency. In this case, it is obvious that the number of matches is either zero or one. One can then split equation 3.35 into two cases. First, the case where $M = 0$, with one permutation allowed in ζ and λ. Second, the case of $M = 1$, where λ still only has one permutation, due to its catalogue having length one. This reduces the triple sum to a sum over γ, each star being the counterpart in turn, which is the sum over j in equations 6 and 7 of Naylor, Broos, and Feigelson (2013). Equivalently, starting from equations 3.32 and 3.33 it is possible to recover equations 6 and 7 of Naylor, Broos, and Feigelson (2013) by forcing the number of elements over which i is iterated to be one, which removes the $i \neq s$ product, and reduces the sum over s and t to just one over t.

This reduction in dimensionality is possible if and only if the separation between stars in catalogue ϕ is much greater than the average radial offset of their counterparts in catalogue γ. This means there is no overlap and no two catalogue ϕ stars can possibly have the same star in catalogue γ within a given radial offset of both stars. Additionally, Naylor, Broos, and Feigelson (2013) made the assumption that the two catalogues' magnitudes are independent of each other, and thus $c(m_k, m_l) = c(m_k)\, c(m_l)$. Finally, two implicit

assumptions were made. The first is that $c(m_l) = f_\phi(m_l)$. Second, the assumption was made that catalogue ϕ is complete, meaning that the symmetrisation of the counterpart magnitude probability density in Section 3.5 is not required, effectively setting $p_\phi = 1$.

To introduce the concept of X into my equations (see table 1 of Naylor, Broos, and Feigelson, 2013) I define it as the fraction of stars with counterparts in catalogue ϕ,

$$X = \frac{N_\mathrm{c}}{N_\phi + N_\mathrm{c}}. \tag{3.58}$$

Rearranging the terms,

$$\frac{N_\mathrm{c}}{N_\phi} = \frac{X}{1 - X}. \tag{3.59}$$

The correct ratios found in $p(H_a|D)$ can now be reproduced. To do so, I start with my original equations 3.32 and 3.33, restated in their compact notation (Section 3.3.3.1) as

$$P(H_a|D) = \frac{N_\mathrm{c} G^{kl}_{\gamma\phi} c^{kl}_{\gamma\phi} \prod\limits_{i\neq k} N_\gamma f^i_\gamma \prod\limits_{j\neq l} N_\phi f^j_\phi}{\prod\limits_i N_\gamma f^i_\gamma \prod\limits_j N_\phi f^j_\phi + \sum\limits_s \sum\limits_t N_\mathrm{c} G^{st}_{\gamma\phi} c^{st}_{\gamma\phi} \prod\limits_{i\neq s} N_\gamma f^i_\gamma \prod\limits_{j\neq t} N_\phi f^j_\phi}, \tag{3.60}$$

and

$$P(H_0|D) = \frac{\prod\limits_i N_\gamma f^i_\gamma \prod\limits_j N_\phi f^j_\phi}{\prod\limits_i N_\gamma f^i_\gamma \prod\limits_j N_\phi f^j_\phi + \sum\limits_s \sum\limits_t N_\mathrm{c} G^{st}_{\gamma\phi} c^{st}_{\gamma\phi} \prod\limits_{i\neq s} N_\gamma f^i_\gamma \prod\limits_{j\neq t} N_\phi f^j_\phi}. \tag{3.61}$$

First the length of catalogue γ is set to one, which removes the product $\prod\limits_{i\neq k} N_\gamma f^i_\gamma$ and reduces the product $\prod\limits_i N_\gamma f^i_\gamma$ to $N_\gamma f^k_\gamma$. Any terms containing $\prod\limits_{j\neq l} N_\phi f^j_\phi$ in equations 3.60 and 3.61 are multiplied and divided by $N_\phi f^l_\phi$, for both l and t.

Switching back to the full notation, all terms in equations 3.60 and 3.61 are divided by

$$N_\gamma f_\gamma(m_k) \prod_j N_\phi f_\phi(m_j). \tag{3.62}$$

This gives

$$P(H_a|D) = \frac{\frac{N_c G(\Delta x_{kl},\Delta y_{kl}) c(m_k,m_l)}{N_\gamma f_\gamma(m_k) N_\phi f_\phi(m_l)}}{1 + \sum_l \frac{N_c G(\Delta x_{kl},\Delta y_{kl}) c(m_k,m_l)}{N_\gamma f_\gamma(m_k) N_\phi f_\phi(m_l)}} \tag{3.63}$$

and

$$P(H_0|D) = \frac{1}{1 + \sum_l \frac{N_c G(\Delta x_{kl},\Delta y_{kl}) c(m_k,m_l)}{N_\gamma f_\gamma(m_k) N_\phi f_\phi(m_l)}}, \tag{3.64}$$

re-introducing the likelihood ratio to my probabilities.

Therefore, after splitting $c(m_k, m_l)$ into $c(m_k)c(m_l)$; cancelling $c(m_l)$ and $f_\phi(m_l)$, assumed to be equivalent; substituting for equation 3.59; and multiplying by $1 - X$, I recover equations 6 and 7 of Naylor, Broos, and Feigelson (2013),

$$P(H_a|D) = \frac{\frac{X g(\Delta x,\Delta y)}{N_\gamma} \frac{c(m_a)}{f_\gamma(m_a)}}{1 - X + \sum_\alpha \frac{X g(\Delta x,\Delta y)}{N_\gamma} \frac{c(m_\alpha)}{f_\gamma(m_\alpha)}} \tag{3.65}$$

and

$$P(H_0|D) = \frac{1 - X}{1 - X + \sum_\alpha \frac{X g(\Delta x,\Delta y)}{N_\gamma} \frac{c(m_\alpha)}{f_\gamma(m_\alpha)}}. \tag{3.66}$$

Note that the g term of Naylor, Broos, and Feigelson (2013) is my G, as they add a systematic uncertainty to their X-ray uncertainties, believed to reflect the infrared uncertainties, and thus it is a convolution of two Gaussians.

While the appendix derivation of Naylor, Broos, and Feigelson (2013) required $P(H_0) = 1 - X$ and $P\left(\widetilde{H}_0\right) = X$, my new derivation contains these implicitly as the ratio of counterparts per unit area to unmatched stars per unit area. I therefore have indifferent priors, assuming a flat prior across all hypotheses. This is required in my formalism due to the extension to a symmetric handling of stars in both catalogues, as well as the extension to multiple potential counterparts in each catalogue. The number densities of matched and unmatched objects can only be considered as simple Bayesian priors in the case where the information of only one catalogue is used, for one potential counterpart. However, the end result is identical, and the equations correctly reduce to their original forms in various limits.

3.9 Conclusions

I have developed a new symmetric method for assigning stars between two catalogues as either counterparts, or unrelated and unmatched stars. I use the extra information gained from the measured photometric magnitudes of the stars to more accurately accept or reject star pairings. My more general formalism for the astrometric probability formally describes the handling of astrometric uncertainties in an equal fashion. It also allows for a more general inclusion of systematic astrometric effects such as proper motion or contamination caused by stellar crowding. I have also expanded the treatment of photometric probabilities to a two-directional treatment, asking the probability of a star having the detected magnitudes of both objects. This new method also allows for the possibility of multiple choices of counterpart for stars in each catalogue. Additionally, I showed how to extend the method to multiple catalogues.

I tested the method on three catalogues: IPHAS, 2MASS, and *Gaia*. I showed that the method correctly returns counterparts in the expected regimes of shared dynamical range between two given catalogues. When compared to a 3 arcsecond nearest neighbour-based match, I successfully return more unassigned, unmatched objects at very bright and very faint magnitudes, outside of the dynamical range of the opposing catalogue. I

also show that the method works when applied to two catalogues of similar astrometric precision, with a truly symmetric handling of the assigning of counterparts between catalogues. In all catalogue match cases, and in all brightness regimes, the inclusion of the photometric likelihoods allowed for a more robust determination of the corresponding objects between catalogues, providing on average a factor 10 improvement to the Bayes' factor. This provides the ability to break nearest neighbour and pure astrometric probability match degeneracies.

The nature of the method gives the flexibility to choose a probability above which to accept counterparts, allowing for the option of only selecting very likely joins between catalogues, giving the confidence in the resulting SEDs.

Chapter 4

Including the Effects of Crowding in the Cross-Matching of Photometric Catalogues

When you do things right, people won't be sure you've done anything at all.

— Cosmic Entity, Futurama (2002)

4.1 Introduction

There are cases when it is necessary to be flexible in the description of the astrometric uncertainty functions (AUFs) used in the probability-based matching process. For *Wide-field Infrared Survey Explorer* (*WISE*; Wright et al., 2010), I concluded in Chapter 2 that confusion caused by the crowding of faint contaminant stars (caused by effects of finite pixel size or point-spread function width) was a significant source of systematics in the AUFs. The undetected contaminating stars inside a bright object's point-spread function (PSF) lead to an AUF with a long, non-Gaussian tail. These perturbations act on length scales much greater than the typical perturbation due to non-zero proper motion, perhaps the most common additional cause of systematic perturbation. In turn, some separations

between likely counterparts become much larger than previously assumed, even after accounting for smaller scale perturbations such as proper motion. Therefore, when considering a catalogue with significant crowding, like *WISE*, the effect of contaminants on the measured positions cannot be ignored. If ignored, the non-Gaussian tails to the AUFs will reduce the astrometric likelihoods to a sufficient level to result in probability-based matches that return significantly fewer counterparts than a simple nearest neighbour-based match.

The matching process described in Chapter 3 combines a flexible formalism of the AUFs describing the detections in each photometric catalogue with the inclusion of the photometric information from both catalogues. This allows for the assignment or rejection of counterpart pairings on both astrometric and photometric probabilities, providing robust pairings with a low false match rate. I adapt this method to include the perturbation from faint contaminant stars as described in Chapter 2. Additionally, I more generally describe the procedure for the implementation of the effects of contaminant perturbation in the AUF. This method allows for flexible modelling of the effect of fainter sources blended into the PSF of any photometric catalogue.

The layout of the chapter is as follows. Section 4.2 describes the probability-based matching of *WISE* and *Gaia* Data Release 1 (DR1; Gaia Collaboration et al., 2016b; Gaia Collaboration et al., 2016a) in the case where contamination is not taken into account. Section 4.3 details how to correct the AUF of a probability-based matching method empirically to include the effects of crowding, and applies the method to the *Gaia*-*WISE* matching case. Section 4.4 includes analysis of test *Gaia*-*WISE* cases, motivation on the merits of applying the matching process to a wider area, comparison with previous results, and a discussion of the photometric effects of crowding. Here I show that the additional matches that are astrometrically perturbed enough to be missed by a Gaussian probability-based match are flux contaminated by an average of 27%. I also compare the *WISE* matches to *Spitzer* (Werner et al., 2004), showing that the higher angular resolution of *Spitzer* sometimes allows for the resolving of the hidden *WISE* contaminants. Section

4.5 provides a brief discussion of several implications these results have, highlights a few minor caveats, and discusses extensions to the methodology. Concluding remarks are then given in Section 4.6. Table 1.1 defines symbol usage in the chapter.

4.2 The Gaussian Astrometric Uncertainty Function

Before the significance of the inclusion of perturbations in the description of the AUFs can be quantified, first the matches obtained without their consideration must be discussed. Therefore the first choice of AUF should be the most obvious, the assumption made most often in probability-based matching: that the probability of two detections of a source being at a given separation is entirely described by a Gaussian. In this section I will describe the results of matching *WISE* to *Gaia* in a crowded region of the Galactic plane under the assumption of a purely Gaussian AUF.

4.2.1 Constructing the Gaussian AUF

When using a probability-based matching method, the astrometric probability density function (PDF) is usually assumed to be a two-dimensional zero-centered Gaussian with covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\delta \\ \rho\sigma_\alpha\sigma_\delta & \sigma_\delta^2 \end{pmatrix}, \tag{4.1}$$

where σ_α and σ_δ are the convolved *Gaia*-*WISE* uncertainties in the two orthogonal sky directions (Right Ascension and Declination, respectively) and ρ is the correlation between the two. G is then

$$G(\Delta\alpha, \Delta\delta) = \frac{\exp\left(-\frac{1}{2\sqrt{1-\rho^2}}\left(\frac{(\Delta\alpha)^2}{\sigma_\alpha^2} + \frac{(\Delta\delta)^2}{\sigma_\delta^2} - \frac{2\rho\Delta\alpha\Delta\delta}{\sigma_\alpha\sigma_\delta}\right)\right)}{2\pi\sigma_\alpha\sigma_\delta\sqrt{1-\rho^2}}, \tag{4.2}$$

where $\Delta\alpha$ and $\Delta\delta$ are the orthogonal sky axis offsets between the respective *Gaia* objects and *WISE* sources, including the cosine of the declination which converts the right ascension separations entirely to seconds of arc.

4.2.2 The Effects of the Gaussian AUF on *Gaia*-*WISE* Matches

To test the effect the AUF has on the resulting pairings, I matched *WISE* stars against *Gaia* stars. For a 42 square degree region of the Galactic plane, $131 \leq l \leq 138$, $-3 \leq b \leq 3$, the catalogues were filtered for poor quality, non-stellarity and non-detections as described in Table 1.2. The probability-based matching process laid out in Chapter 3 was used. In all cases it is assumed that G (the convolution of each source's AUF), and any defining merging/cutout radii $\mathcal{R}_Y$ (the circle radius inside which the integral of G is equal to Y), are Gaussian. These functions are described in further detail in Chapter 3. This assumption is also used when using the photometric information available in the catalogues to construct c and f and evaluate the photometric probabilities, also detailed in Chapter 3.

The results of this cross-match are shown in Figure 4.1, for matches with a probability $P \geq 0.5$ (see Section 3.3.3 for details on how the match probabilities are calculated). The counterparts the cross-matching process returns have magnitudes which lie in a sensible region of the $G - W1$ magnitude-magnitude plane (main panel, Figure 4.1). As expected, the density of matches increases towards fainter G and $W1$ magnitudes, with *Gaia* magnitudes typically 1-4 magnitudes fainter than the *WISE* passbands. The dwarf-giant separation is also recovered towards brighter magnitudes ($W1 \leq 12$).

However, as shown by the side panels of Figure 4.1, the assumption that the positional uncertainties are described by a Gaussian results in only 52% of the matches that were returned using a 3 arcsecond nearest neighbour-based matching procedure. Assuming a *Gaia* stellar density of 2×10^4 deg^{-2} in the area of the Galactic plane in question (Section 2.4.2.2) a false match rate on the order of 4% is found. These additional nearest neighbour-based matches cannot be entirely explained as false matches, the removal of approximately 4% of the nearest neighbour matches is expected yet 48% of these matches are rejected in

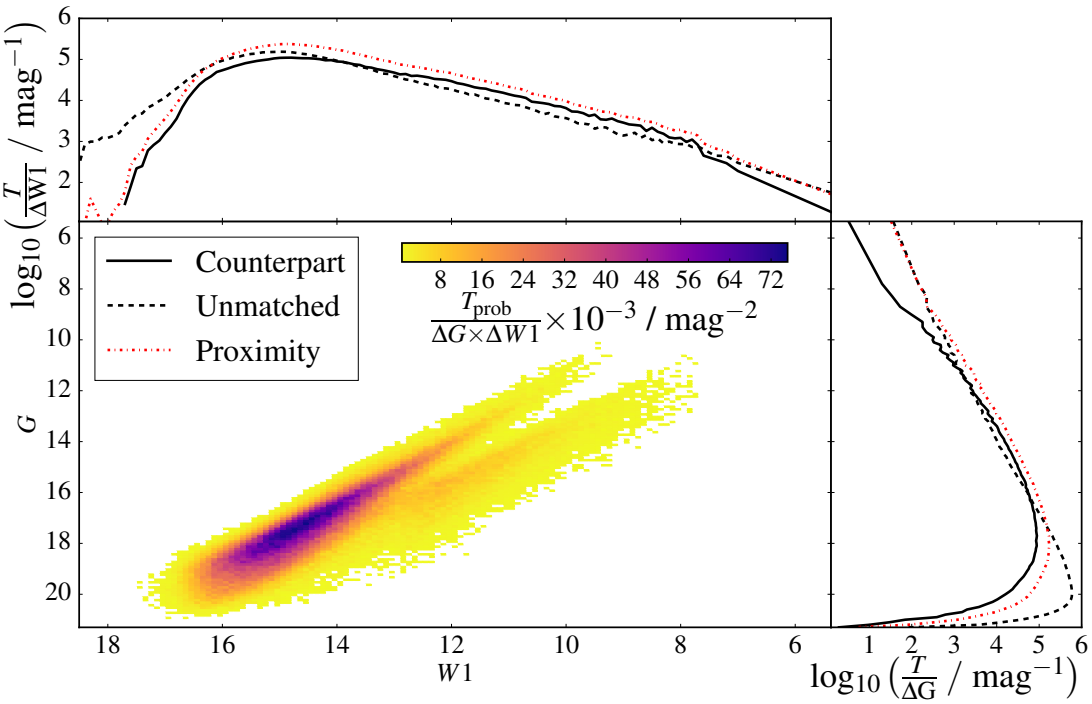


Figure 4.1: The number density of matched objects between *WISE* and *Gaia* for a 42 square degree region of the Galactic plane. The main panel shows a 2D histogram of the number density of objects, for objects with both G and $W1$ detections as a function of G and $W1$. The two inset panels show the number density of objects as functions of G or $W1$ magnitude alone. The inset panels show the results of 3 arcsecond nearest neighbour matches in a red dash-dotted line, probability-based counterparts in a solid black line, and probability-based unmatched "field" stars in a black dashed line. Using a Gaussian to represent the AUF results in matches for 52% of the nearest neighbour matches. Only bins with densities $\geq 500\mathrm{mag}^{-2}$ are displayed in the main panel.

the probability-based match. This order-of-magnitude increase in rejection rate must have a different explanation.

Considering the likelihood ratios (e.g., Sutherland and Saunders, 1992) of the astrometric and photometric halves of the equations used in the probability-based matches (see Section 3.6.4.1 for more details) shows the reason for the loss of these nearest neighbour matches. The photometric likelihood ratio, η, is defined as the logarithm of the ratio of the counterpart probability density, c, to the likelihood of the two unmatched densities, $f_\gamma \cdot f_\phi$. Equivalently, the astrometric likelihood ratio, ξ, logarithmically balances the astrometric counterpart probability density, $N_c G$, with the probability density of two unrelated objects, $N_\gamma \cdot N_\phi$.

As shown in Figure 4.2, the majority of the objects matched return both a high astrometric ($\xi \geq 0$) and photometric ($\eta \geq 0$) likelihood ratio; they are more likely than not to be matched on both spatial and magnitude grounds. At very high matched object astrometric likelihood ratios ($\xi \geq 2$) the photometric likelihood is very high as well. At lower (albeit still more than equal probability) astrometric likelihood ratios the photometric likelihood ratio of these matched objects plateaus at $\eta \simeq 0.3$.

Also shown in Figure 4.2 are the likelihood ratios for any pairs that are nearest neighbour matched within 3 arcseconds but not returned as a pair by the probability-based match. These are all below the equal likelihood ratio line, defined as being $\xi + \eta = 0$. However, they still follow $\eta \simeq 0.3$, implying a photometric likelihood ratio higher than equal chance, and no lower on average than the returned probabilistic matches. Therefore, the matches are failing to be returned due to their astrometric likelihood ratio, which rapidly decreases to several orders of magnitude below equal likelihood.

4.3 The Empirical Astrometric Uncertainty Function

Since the probability-based matches are being lost on purely astrometric arguments, the definition of G must be reconsidered to correct for the missing $\simeq 50\%$ of the nearest neighbour counterparts. To achieve this, empirical AUFs can be constructed, based on

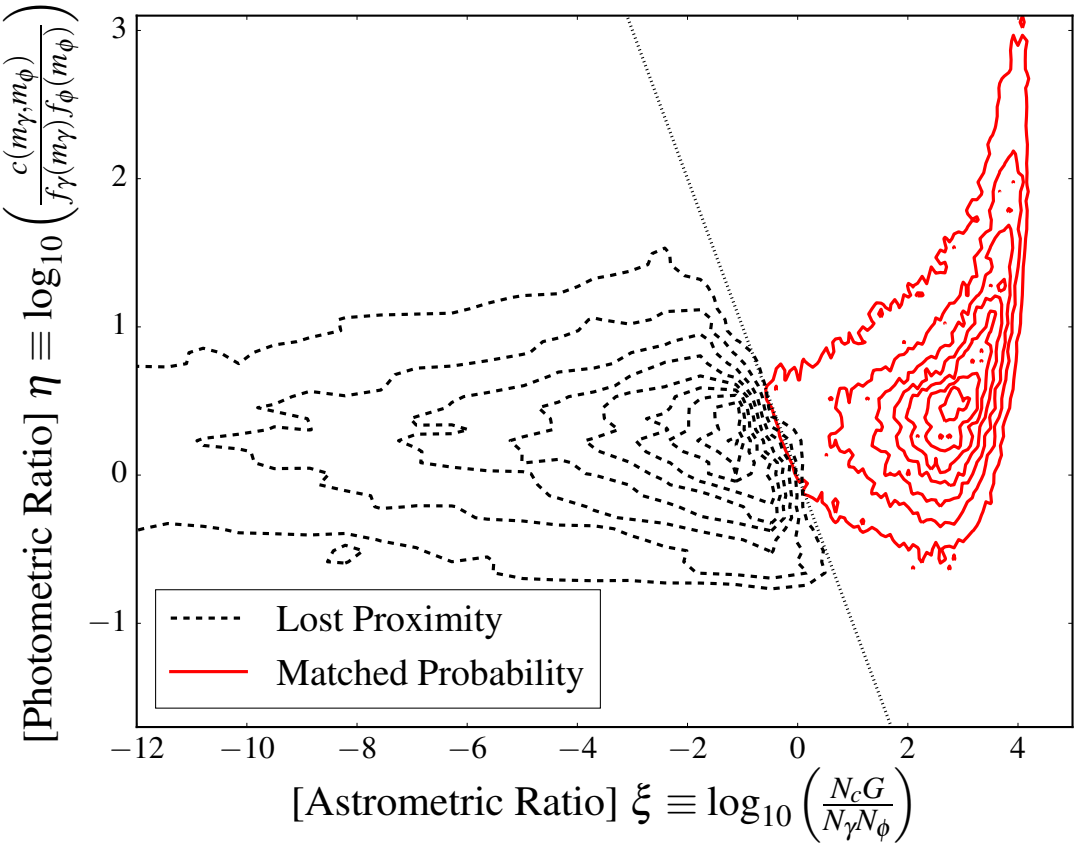


Figure 4.2: The photometric and astrometric likelihood ratios of *Gaia* matches for a 42 square degree region of the Galactic plane, under the assumption that the distribution of separations follows a Gaussian distribution. The density of objects returned as pairs by the probability-based match is shown as red solid contours. Objects paired by a 3 arcsecond nearest neighbour match but returned as unrelated in the probability-based match are shown as the dashed black contours. The lost matches are below the equal probability dotted line, $\xi + \eta = 0$, but at a lower astrometric likelihood ratio, with a roughly constant photometric likelihood ratio $\eta \simeq 0.3$. This implies the pairings are more likely than not based on photometry arguments, but are lost due to the assumptions made about G.

the distribution of separations for a given area of the sky. This allows for varying levels of crowding seen at varying longitudes and latitudes throughout the sky to be taken into account.

4.3.1 Constructing the Empirical AUF

To model the AUF of perturbed stars, some simple numerical models are required, as discussed in Chapter 2. In these the effects of a given stellar density are simulated, recording the positions of stars inside the bright, central star's PSF, including the effects of stars well below the completeness limit of a given survey. This distribution is then combined with the intrinsic positional uncertainty of the given star. The resulting PDF, for the offset of the star from its true position, is the perturbed-star AUF required for the given star.

First the distribution of physical perturbations of a star in the stellar field in question must be obtained. For this example I assume, following Chapter 2, that the stellar density of objects as a function of magnitude in a given filter, D, follows a geometric series. This density gives the number of stars per unit magnitude per unit area, $D = Nz^m$, at magnitude m. Here N is the stellar density (per unit magnitude per unit area) of the field at zeroth magnitude, and z is the geometric scaling factor which dictates the rate of increase of the stellar density with decreasing brightness. In addition, since only stars within our PSF circle are of interest, a term for the circle area must be included, giving an effective "magnitude density" at the magnitude of the star,

$$B = Nz^m \pi R^2, \tag{4.3}$$

where R is the radius of the PSF circle. For this radius I use the Rayleigh criterion (Rayleigh, 1880) of the telescope

$$R = 1.185 \times \mathrm{FWHM} \tag{4.4}$$

as described by Airy (1835), where FWHM is the full width at half maximum of the telescope Airy disk or atmospheric seeing (typically on the order of one arcsecond) PSF, whichever is larger. To build up a sample PSF contamination the chance of a contaminant star of given magnitude offset (i.e., with a certain flux ratio) Δm relative to the bright central source (of magnitude m) being in the PSF circle must be evaluated. The average number of stars of each faint magnitude slice $m + \Delta m$ is given by integrating the magnitude density across the bin width (dm),

$$P_B(m + \Delta m) = \int_{m+\Delta m}^{m+\Delta m+\mathrm{d}m} B(m')\,\mathrm{d}m' = \int_{m+\Delta m}^{m+\Delta m+\mathrm{d}m} N z^{m'} \pi R^2\,\mathrm{d}m' = \frac{N z^{m+\Delta m}\left(z^{\mathrm{d}m} - 1\right)}{\log(z)} \times \pi R^2. \tag{4.5}$$

Using this typical source count, the expected number of stars in the PSF at this magnitude slice are drawn from a Poissonian distribution. If non-zero, these stars are randomly distributed in θ and r^2 space (to account for the additional r term in the unit circle area term). These are then converted to Cartesian coordinates. This is repeated for $\Delta m = 0$ to $\Delta m = 10$ in steps of $\mathrm{d}m = 0.025$. Once all magnitude slices have had stars randomly drawn and distributed, the flux-weighted average x and y positions are recorded, and converted back to a radius as $r = \sqrt{x^2 + y^2}$.

This sampling of contaminant star brightnesses and radial offsets is repeated for a million unique test PSFs, each time stepping through Δm. This results in a distribution of offsets, which is then converted to a PDF. This perturbation PDF, h_{offsets}, is then convolved with the Gaussian, h_{pure}, of the intrinsic positional uncertainty of the central star, σ_{pure}, to produce a numerical AUF for a given stellar density, brightness, and positional uncertainty. Mathematically, this is given by

$$h_{\mathrm{tot}} = h_{\mathrm{pure}} * h_{\mathrm{offsets}}. \tag{4.6}$$

An example of such an AUF is shown in Figure 4.3. The intrinsic Gaussian AUF (blue solid line) is convolved with the distribution of perturbations (black histogram), resulting

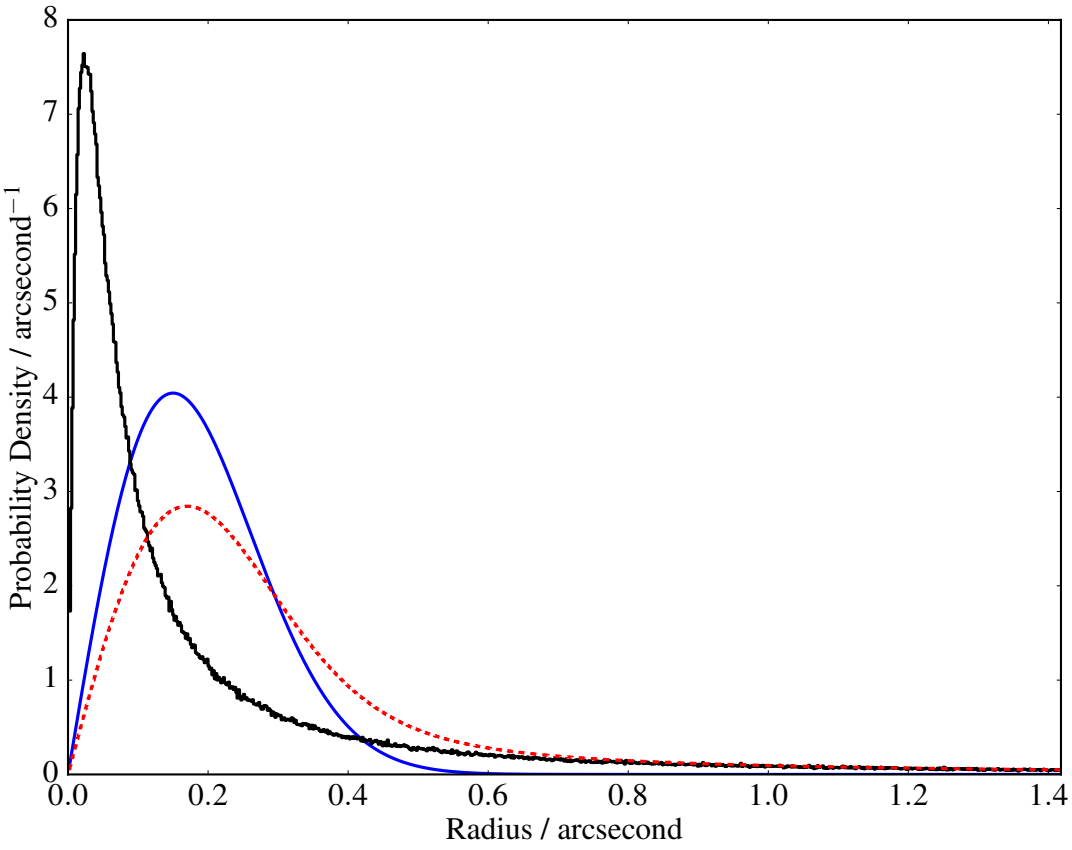


Figure 4.3: An example numerical AUF, for $N = 0.5\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$, $z = 2$, $\sigma_\alpha = 0.15$ arcsecond, $m = 12$. In this case the contaminating stars are drawn from a distribution from equal brightness down to a magnitude difference $\Delta m_{\mathrm{max}} = 10$, or a flux ratio of 0.0001. The black histogram shows the distribution of central star perturbations from the origin, caused by the flux-weighted average positions of the contaminating stars. The blue solid line shows the "pure" Gaussian from which the measured position would be naively drawn, represented here by a Rayleigh distribution, the transformation of a two-dimensional Gaussian to one-dimensional radial coordinates. The red dashed line shows the convolution of the two, giving the resulting AUF.

in an empirical AUF (red dashed line) that includes the effects of the blending of faint contaminant stars into the PSF of the central source on its astrometric position.

In cases of very low crowding, either through high angular resolution and thus small R, low source densities and thus low N, or through bright central magnitude and thus low z^m, the central offset will tend to zero in most numerical simulations. In these cases h_{offsets} reduces, effectively, to a delta function ($\delta_{\mathrm{offsets}}$). For these low crowding cases the AUF is simply the intrinsic Gaussian AUF in the absence of any contamination,

$$h_{\mathrm{tot}} = h_{\mathrm{pure}} * \delta_{\mathrm{offsets}} = h_{\mathrm{pure}}. \tag{4.7}$$

The AUF is then simply a convolution of the "offset" AUF component with the "pure" AUF component, regardless of the levels of contamination suffered by any individual source.

4.3.2 The Dependences of Empirical AUF Construction

The parameterisation of the level to which contaminant stars affect the astrometric position, B (see equation 4.3), is dependent on three further parameters: first, the brightness of the central source, m; second, the overall source density in the region of sky in question, N; and third, a description of the increase in source counts with increasing magnitude, z. The effects of this parameterisation must therefore be explored before the empirical AUFs can be constructed across a large area of the Galactic plane.

4.3.2.1 The Dependence of N and z on l and b

First the decision must be made whether z should be described as a function of l and b. Dense regions of gas (e.g., molecular clouds) will, in theory, cause differential extinction preferentially extincting more distant and fainter stars. However, if the assumption of a constant geometric scaling is made (e.g., Chang, Ko, and Peng, 2010), it would greatly simplify the creation of empirical AUFs, allowing for the total source density to simply scale through the choice of N.

To test this, I initially fitted the differential source counts in a small region of Galactic plane, $134 \leq l \leq 134.2$, $2 \leq b \leq 2.2$, small enough that there should be limited effects from differing source densities across star forming regions, using Nz^m. I found $z = 1.978$. As this value is very close to 2, I tentatively re-adopt $z = 2$ as the canonical geometric scaling law, but first must ensure that this value is appropriate across a variety of differential crowdings.

The inset panel of Figure 4.4 shows a small region of the Galactic plane, $132.5 \leq l \leq 134.25$, $0 \leq b \leq 0.8$. Two smaller regions of interest are selected, with different source counts and column densities, represented here through the proxy of ^{12}CO integrated brightness temperature, using the FCRAO OGS survey (Heyer et al., 1998). Shown in

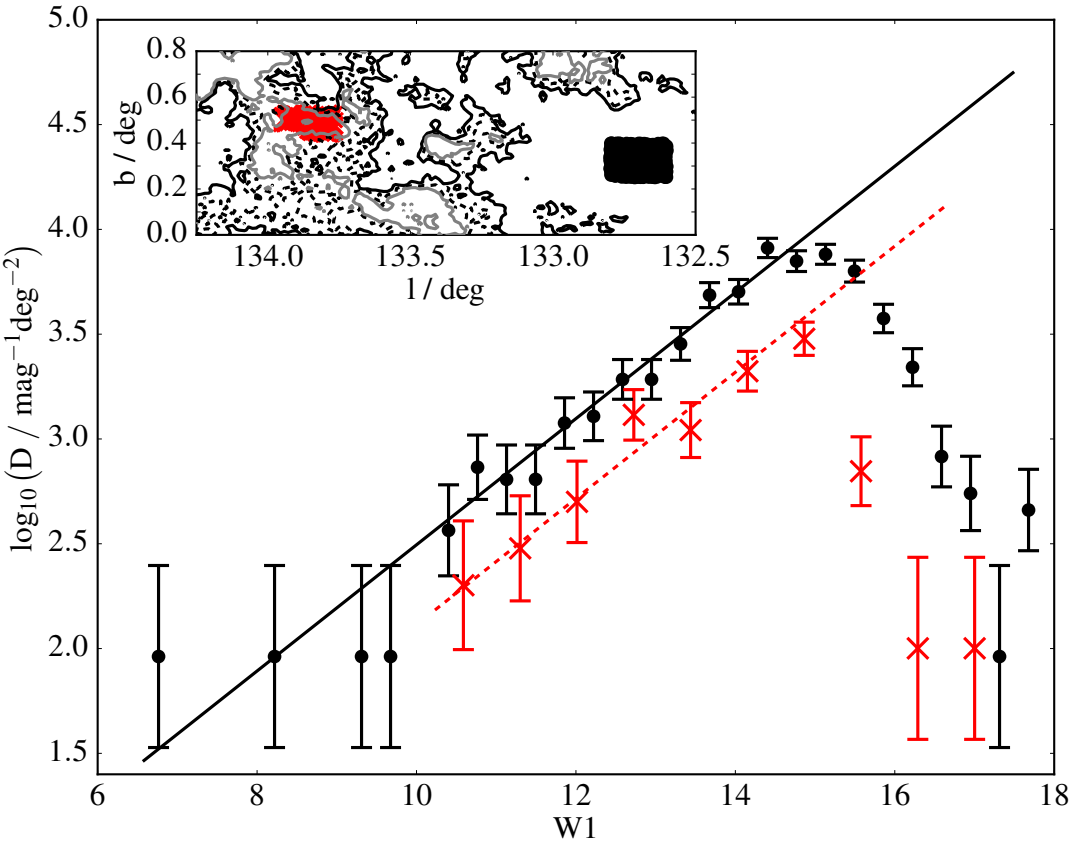


Figure 4.4: The effect of differing stellar densities on source counts. Inset panel shows the spatial distribution of two small patches of *WISE* stars. Contours denote levels of integrated ^{12}CO brightness temperature, 7.23/13.45/19.87 K km s^{-1} respectively, used here as a proxy for column density. Black dots show stars in a region of low integrated brightness temperature, while red crosses mark out a separate region, this time in a higher column density. Main panel shows differential source counts for the two regions. Red dashed and solid black lines represent the best fits to the two datasets respectively, $D = Nz^m$, assuming a geometric scaling law $z = 2$. The two fits have values of $N = 0.127\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$ and $N = 0.304\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$, respectively.

black are stars in a region with low column density, whereas the red data are stars in the line of sight of a molecular cloud, affecting the differential source counts.

The differential source counts for the two regions are shown in the main panel of Figure 4.4, with the best fits to the data, assuming a scaling law of $z = 2$. In both cases the scaling law fits well, with a simple reduction in N for the region of higher column density. I find this relationship fits well across multiple photometric catalogues with differing spatial resolution and wavelength coverage, and therefore suggest $z = 2$ as the invariant bright geometric scaling law across all catalogues and sky positions (see Section 4.3.2.2 for further discussion). However, the intrinsic source density does vary with sky position, and N is still parameterised by the local sky density. This is highlighted in Figure 4.5, for two sets of *WISE* stars of $W1 = 14.97 - 15.03$ and $\sigma_\alpha = 0.06 - 0.12$ arcsecond matched to *Gaia* sources. Neither distribution of separations fits the Gaussian AUF (black dotted line), but both provide excellent fits to their respective fully parameterised AUFs. The key difference between the two, driving the level to which the non-Gaussian wings affect the AUF, is their local normalisation density N. The black data points and solid line are the *Gaia*-*WISE* separations ($131 \leq l \leq 138$, $-3 \leq b \leq 3$) and AUF respectively for stars of the fixed magnitude and uncertainty, but with $N = 0.253 - 0.263\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$. The red data points and line are the *Gaia*-*WISE* separations ($180 \leq l \leq 200$, $8 \leq b \leq 10$) and AUF, for stars with $N = 0.102 - 0.112\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$. The higher overall source density leads, for similar sets of matches – fixed *WISE* magnitude, astrometric precision, and in similar regions in the Galactic plane – to a more perturbed AUF in more crowded local regions of the Galaxy.

4.3.2.2 The Dependence of Differential Source Counts on Central Star Brightness

The assumption was made in Section 4.3.1 that differential source counts as parameterised in Chapter 2 can be extrapolated below the completeness limit of a given survey. However, as shown by, e.g., Bahcall and Soneira (1980), there is a decrease in the count rate of the very faintest objects. This effect has several sources; one of the primary causes of the

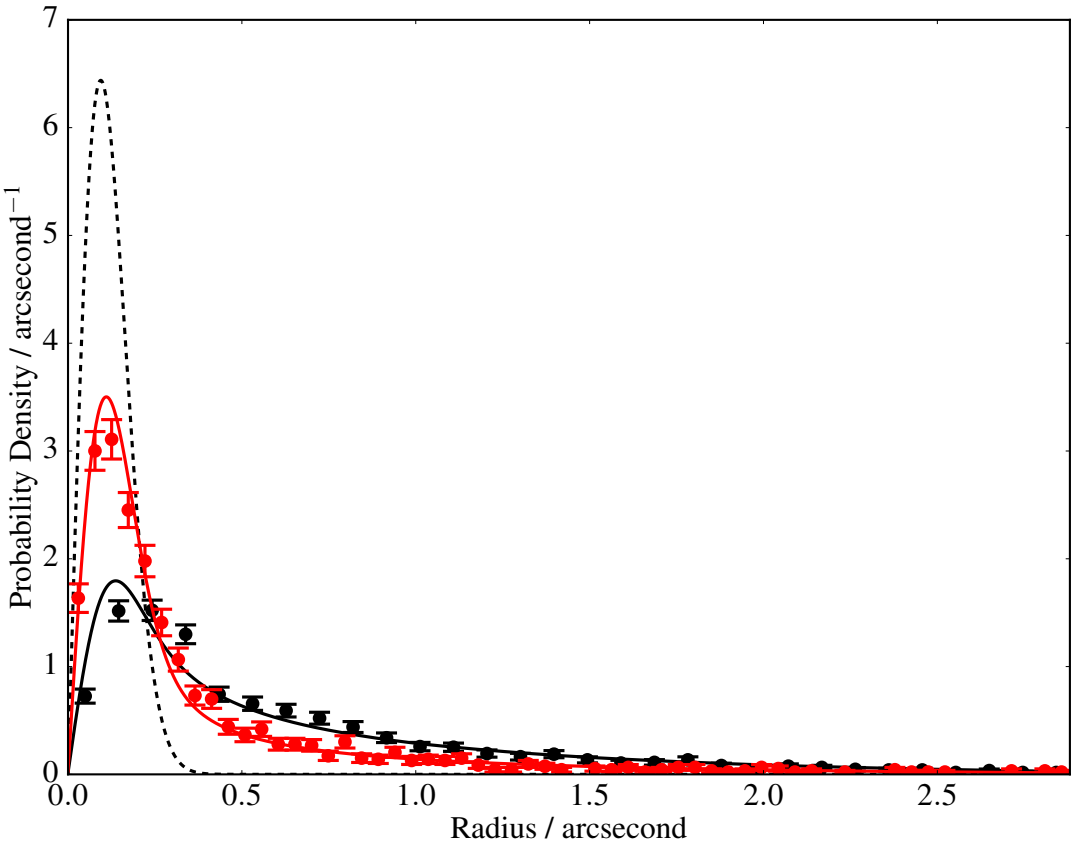


Figure 4.5: The effect of local density on the AUF, for *WISE* stars in the Galactic plane matched to *Gaia* sources. In both examples, stars of $W1 = 15$ and $\sigma_\alpha = 0.09$ arcsecond are selected, with the dotted black line representing a Rayleigh distribution of the given uncertainty. The black error bars show the distribution of separations between *Gaia* stars and *WISE* stars of the given magnitude and astrometric uncertainty and local normalisation density $N = 0.258\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$, with the solid black line showing the AUF for the given magnitude, astrometric uncertainty and normalisation density. Similarly, the red error bars show the distribution of *Gaia-WISE* separations for *WISE* stars of the given magnitude and astrometric uncertainty with $N = 0.107\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$. The higher overall normalisation density results in an increased crowding, and an AUF with larger non-Gaussian wings.

decrease in sources at these magnitudes is the edge of the Galaxy, beyond which stellar densities are much diminished. This turnover means that the previous extrapolation of count rates for a central star of e.g., $W1 = 17$ down a further 10 magnitudes would lead to unphysical contamination fractions; this effect is discussed further in Section 4.5.3. The differential source count model must therefore be re-parameterised to account for this issue at faint magnitudes.

To analyse the differential source counts below the *WISE* completeness limit, a TRILEGAL (Girardi et al., 2005) simulation for one square degree of the Galactic plane centered on $l = 133$, $b = 0$ was obtained. The $W1$ differential source count for the region is shown as black error bars in Figure 4.6. Also shown in Figure 4.6 are three red lines, representing a geometric scaling law parameterisation of the source counts. This multiple law parameterisation is defined by a number of scaling laws (in this case $z_1 = 2$, $z_2 = 1.51$, $z_3 = 0.99$) and crossover magnitudes ($m_2 = 16.5$, $m_3 = 21$). I define each subsequent scaling law normalisation beyond the first as being $N_{i+1} = N_i z_i^{m_{i+1}} / z_{i+1}^{m_{i+1}}$, in which the effective differential source counts for each parameterisation is the same at the crossover magnitude. The entire parameterisation therefore depends solely on the initial normalisation density $N_1 \equiv N$. This is still the source density defined by stars at the bright end of the catalogue, typically easily obtained from detected source counts above the survey completeness limit.

While I have shown that a multi-scaling law parameterisation can remain a reasonable approximation to the differential source counts of a catalogue, I choose to no longer describe the differential source counts analytically. For the remainder of this chapter I choose instead to build the synthetic Galactic *Gaia-WISE* match offset distributions using TRILEGAL simulations to parameterise $z(m)$. I therefore update the method described in Section 4.3.1 to utilise the simulated TRILEGAL stellar population. D is created as a histogram of the simulated magnitude distribution (cf. Figure 4.6), and then the expected number of contaminant stars at each magnitude step is calculated as

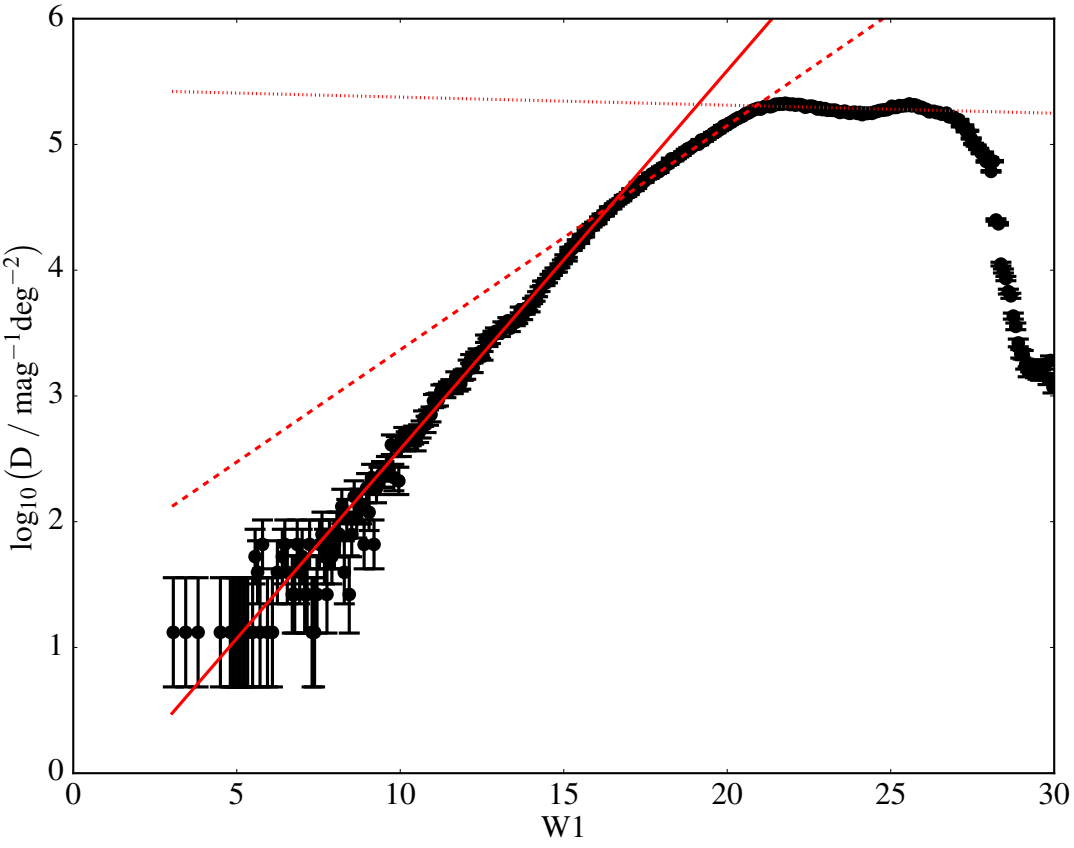


Figure 4.6: TRILEGAL differential source counts for one square degree centered on $l = 133, b = 0$. Also shown are three fits, $D = Nz^m$, to three different parts of the source counts. The solid red line shows the fit to sources $W1 \leq 16.5$, fixed at $z_1 = 2$, for which the best fit value is $N = 0.367\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$. The dashed red line shows the fit to $16.5 \leq W1 \leq 21$, fixed such that the crossover differential source counts are consistent at $W1 = 16.5$, resulting in a $z_2 = 1.51$. The dotted red line shows the fit to $21 \leq W1 \leq 26$, again fixed such that the crossover differential source counts agree at $W1 = 21$, resulting in $z_3 = 0.99$.

$$P_B(m + \Delta m) = \frac{N_{\text{empirical}}}{N_{\text{TRILEGAL}}} D(m + \Delta m) \times \pi R^2 \times \mathrm{d}m, \tag{4.8}$$

where $N_{\text{empirical}}$ is the local bright-magnitude normalising density of the catalogue in question, and N_{TRILEGAL} is the equivalent normalising density of the simulated data. This ratio is a simple correction factor to re-normalise the relative counts to those of the data; the important information the simulated magnitude differential source counts provide is $z(m)$.

This method is used to construct the AUFs used to evaluate the *Gaia*-*WISE* matches in Section 4.4. However, it should be noted that there may be certain cases where such simulations may not be available or relevant, in which case the power law parameterisation may be the preferred choice. This is discussed further in Section 4.5.6 for the case of faint sources out of the plane of the Galaxy, where extragalactic sources dominate the differential source count.

4.3.3 Applying a Empirical AUF to *Gaia*-*WISE* Separations

Now that a complete description of the differential source counts in a given filter has been found, including effects below the catalogues' sensitivity, new AUFs can be constructed. Each empirical AUF is uniquely described by three parameters: N, the geometric scaling normalisation of the bright part of the scaling law; m, the magnitude of the central source; and σ_{pure}, the intrinsic uncertainty of the centroiding of the central source in the absence of crowding. I calculate N by obtaining the number of *WISE* objects in the range $9 \leq W1 \leq 14$ within 15 arcminutes of each *WISE* source, U, and solving the equality

$$U = \int_{m'=9}^{14} N z^{m'} \mathrm{d}m' \times \pi \times (15\,\text{arcminute})^2. \tag{4.9}$$

Once N and m are known each contaminant star magnitude can be incremented through, calculating P_B (cf. equation 4.8, or equation 4.5 for the bright scaling law limiting case) and drawing contaminant stars to place within the PSF. The flux-weighted average of all

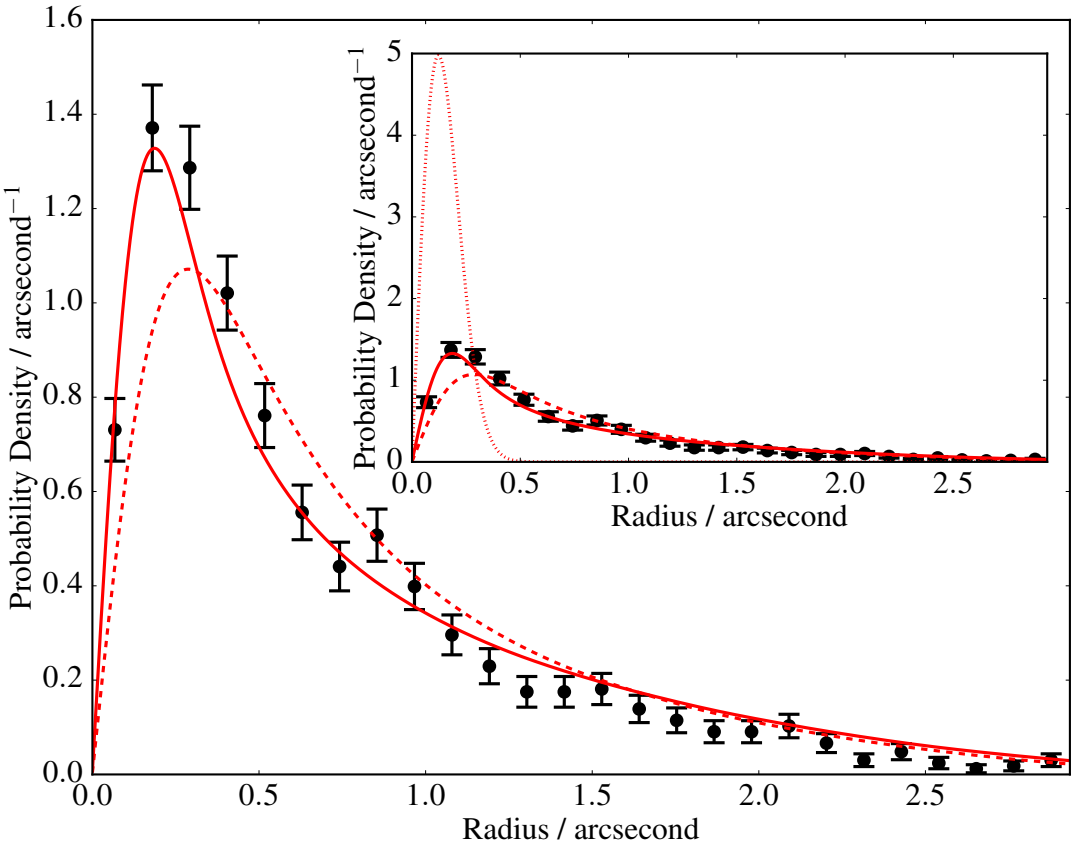


Figure 4.7: The distribution of separations between *WISE* and *Gaia* objects for 42 square degree region of the Galactic plane. Black circles in both main and inset panels show the number density of separations found using a 3 arcsecond nearest neighbour match, for *WISE* objects $N = 0.253 - 0.262\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$, $W1 = 15.47 - 15.03$, $\sigma_\alpha = 0.112 - 0.132$ arcsecond. Solid red lines show the full empirical AUF for these parameters. Dashed red lines show the empirical AUF without the inclusion of the differential source count breaks (Section 4.3.2.2), resulting in a distribution with larger perturbation offsets than seen in the data. Dotted red line in the inset panel shows a purely Gaussian AUF, represented by a Rayleigh distribution with uncertainty 0.122 arcsecond, entirely incompatible with the distribution of *Gaia-WISE* separations.

of the stars in a given PSF can be found, and the process repeated, as described in Section 4.3.1.

An example of the full empirical AUF treatment is shown in Figure 4.7 (solid red line), compared to the 3 arcsecond nearest neighbour matching of *WISE* and *Gaia* objects with $N = 0.258\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$, $W1 = 15.5$, and $\sigma_\alpha = 0.122$ arcsecond. The purely Gaussian AUF (dotted red line, inset panel) is completely incompatible with the separations seen in the data; however, there is good agreement between the empirical AUF and the distribution of separations. The slight discrepancies between the separations and empirical distribution can be explained by a combination of the slight spreads in

values of N, $W1$, and σ_α used to build the *Gaia-WISE* separations. Additionally, I do not include the effects of proper motions in these empirical AUFs, and therefore miss a small additional source of perturbation seen in the separations between sources. The inclusion of this extra perturbation of astrometric positions would slightly broaden the AUFs further. However, as can be seen in Figure 4.7, the epoch differences between *WISE* and *Gaia* cause negligible positional shifts compared to those caused by the crowding of stellar sources (see Chapter 2 for a more detailed discussion). I will discuss the inclusion of the effects of proper motions in AUFs further in Section 4.5.6.1.

4.3.4 Empirical AUF Fitting Summary

I summarise the steps required to compute a given empirical AUF, including the effects of perturbation due to crowding, for a specific star as follows.

1. Determine N, m and σ_{pure}.

2. Create a parameterisation of the differential source magnitude counts for the filter in question.

3. Assign random positions in the PSF to stars for a small magnitude offset range, drawing the number of stars according to Poissonian distribution.

4. Repeat the drawing of stars from the probability distribution for all magnitude offsets, accounting for differential source count variations with magnitude.

5. Using all contaminating stars within the PSF, determine the flux-weighted star position, to find the perturbation offset.

6. Repeat the perturbation offset calculation for a large number of PSFs, creating the offset distribution.

7. Convolve the offset distribution with a pure Gaussian of given uncertainty.

4.3.5 The Effects of the Empirical AUF on *Gaia*-*WISE* Matches

Now that empirical AUFs have been constructed, they can be applied to the same sky region as in Section 4.2. While the two photometric catalogues are still matched using the method laid out in Chapter 3, the empirically constructed AUFs define G. I also define the island cutoff radii, as well as counterpart and "field" star cut out radii as described in Chapter 3, using the new empirical AUFs. Assuming circular symmetry for the AUFs (see Section 4.5.4) simplifies the definition of $\mathcal{R}_Y$ somewhat, however, and it is now defined as

$$\int_0^{\mathcal{R}_Y}\int_0^{2\pi} r\,G(r,\theta)\mathrm{d}\theta\,\mathrm{d}r = 2\pi \times \int_0^{\mathcal{R}_Y} r\,G(r)\,\mathrm{d}r = Y. \tag{4.10}$$

I am more lenient in this chapter than in Chapter 3 with the maximum offset due to the long, non-Gaussian tails, using the largest $\mathcal{R}_{0.99}$ of all *WISE* stars in the matching region in question, slightly less complete than as with a Gaussian G. This slightly lower integral limit is still over an order of magnitude higher than that used in Section 4.2.2, due to the large effect contamination has on the *WISE* positions. The nature of the non-Gaussian tails to the AUF mean that the integrals must be cut at a slightly lower percentile than previously; see Section 4.4.1.1 for discussion of the effect this has on the matches obtained.

Matching the same catalogues as described in Section 4.2.2, the results of using the new PDF for G are shown in Figure 4.8, again accepting only matches with $P \geq 0.5$. The vast majority of the nearest neighbour-based counterparts are now recovered. A reduction of the number of faint (*Gaia* $G \leq 20$, bottom of right inset panel) counterparts is also seen, when compared with the nearest neighbour matches, as expected. However, the objects recovered and rejected at the varying brightnesses in both the *Gaia* and *WISE* passbands require more detailed examination.

The number of objects gained or lost by the probability-based matching process relative to the 3 arcsecond nearest neighbour match can therefore now be considered, as shown in Figure 4.9. The first point of interest is that over much of the area occupied by bright ($W1 \leq 14$) matches there is a rejection of approximately 1-5% of the matches,

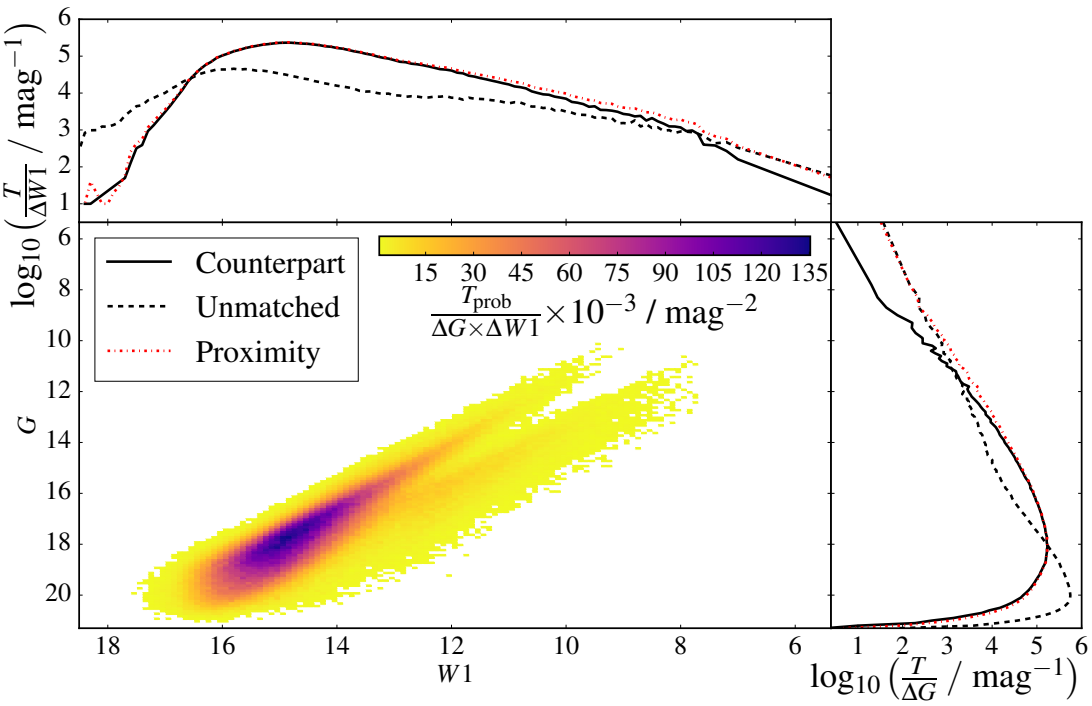


Figure 4.8: The number density of matched objects between *WISE* and *Gaia* using probability-based matching that includes the effect of crowding in the AUF for a 42 square degree region of the Galactic plane. Figure layout and colourbar are the same as Figure 4.1. The empirical *WISE* AUF results in a much more complete counterpart return rate, recovering more counterparts than the nearest neighbour-based match at $G \simeq 18$. It still rejects faint matches $G \geq 20$ as required.

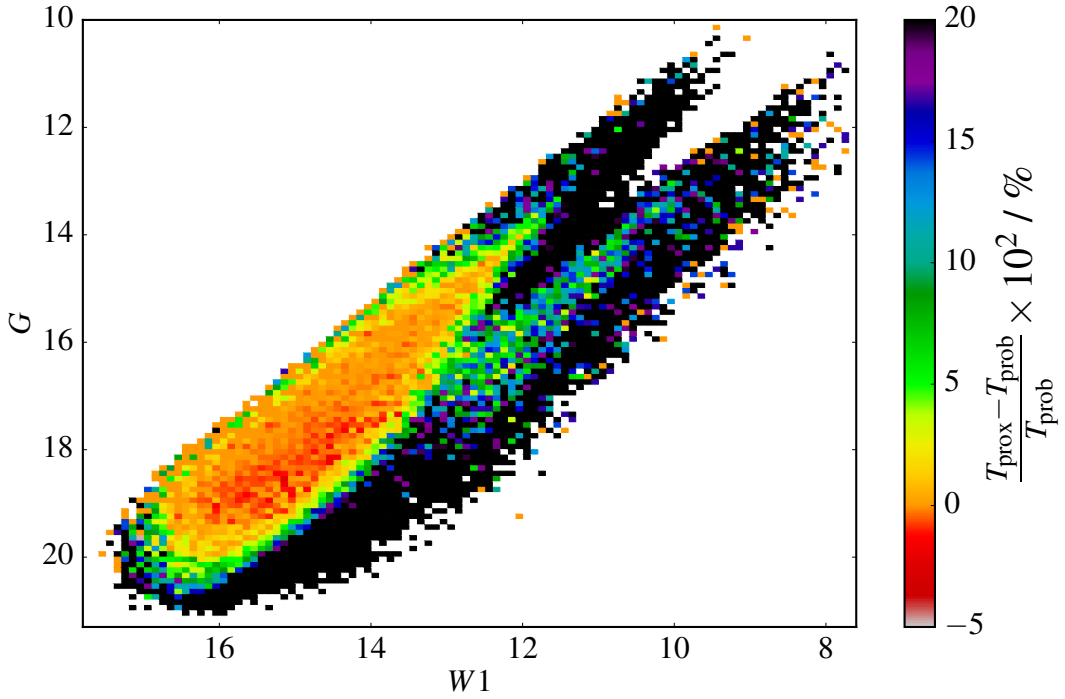


Figure 4.9: The relative difference in the number of objects in a 42 square degree region of the Galactic plane for *Gaia-WISE* objects. The magnitude density criterion is the same as for the main panel of Figure 4.1. However, the colourbar shows the relative difference in probability- and nearest neighbour-based ("proximity") matches. $W1 \leq 14$ there is a constant rejection of a small number of objects in all bins on the order of several percent, consistent with false match chance arguments. However, at $W1 \simeq 15$ there are two areas of importance. First, at $G \simeq 18$ the probability-based matches return additional matches not picked up at a 3 arcsecond nearest neighbour match, suggesting a small number of objects are astrometrically perturbed by >3 arcsecond. Second, at $G \simeq 20$, there is a significant decrease in the number of matches.

similar to the number of false positives (see Section 4.2.2). This indicates that the new AUF is still rejecting false matches, as expected.

At faint magnitudes ($W1 \simeq 15$) there are two distinct regions of the magnitude-magnitude space. The first, at $G \simeq 18$, is an area where extra pairings are picked up by the probability-based matching, which were not picked up by the nearest neighbour match. These are most likely objects which were astrometrically perturbed beyond the nearest neighbour cutoff radius, and therefore unable to be paired in the nearest neighbour match. The contamination at this magnitude is most likely to cause astrometric shifts which result in separations between *WISE* and *Gaia* source detections beyond the 3 arcsecond nearest neighbour match radius (see Chapter 2). However, some of them could also be objects where the pair most favourable was not the closest. These objects would favour brighter,

but further away, matches rather than some fainter, but closer, stars. This can be caused either by the brighter source having a larger absolute distance but smaller Mahalanobis distance, due to its smaller astrometric uncertainties, or by the photometric counterpart likelihood favouring the bright source over the faint object. The second region of interest, at fainter *Gaia* magnitudes ($G \simeq 20$), sees a loss of matches compared with nearest neighbour match for the same *WISE* brightness ($W1 \simeq 15$). These could be the rejected faint nearest neighbour matches for the additional probability-based matches seen at $G \simeq 18$. However, a fraction of these lost, faint *Gaia* matches are *WISE* objects which should match to *Gaia* objects below the sensitivity level of the survey, which are coincidentally near to these objects of $G \simeq 20$ whose corresponding *WISE* object was removed from the catalogue in the process of cleaning poor quality data (see Table 1.2). This issue with incomplete datasets and quality selection can also explain the lack of bright pairings, where objects of $W1 \simeq 7$ should match *Gaia* sources of $G \simeq 11$. Those rejected pairings (dashed lines, inset Figure 4.8 panels) are primarly caused by saturation effects, with *WISE* having a saturation magnitude $W1 \simeq 8$.

The acceptance and rejection of the nearest neighbour matches on probabilistic grounds can be analysed by considering the likelihood ratios once more, shown in Figure 4.10. Most matches are still several orders of magnitude more likely matches than non-matches, based on their astrometry. Additionally, the spread of η values ($\overline{\eta} \simeq 0.6$, $-1 \lesssim \eta \lesssim 2$) is the consistent with the case where the AUF was purely Gaussian. The differences arise when considering those objects rejected as probability-based matches which were nearest neighbour matched at 3 arcseconds. With the empirical G term, the matches which are now lost with respect to the nearest neighbour matches still have $\xi \geq 0$, but are an order of magnitude less likely to be a match to their nearest neighbour, than to be unrelated, photometrically (i.e., $\eta \simeq -1$). This suggests that those objects still not matched to a star positionally close to them when using the empirical AUF are rejected for flux-related reasons. This is in contrast to the Gaussian AUF case (cf. Figure 4.2), where the losses were almost all astrometric. The inclusion of the photometric information

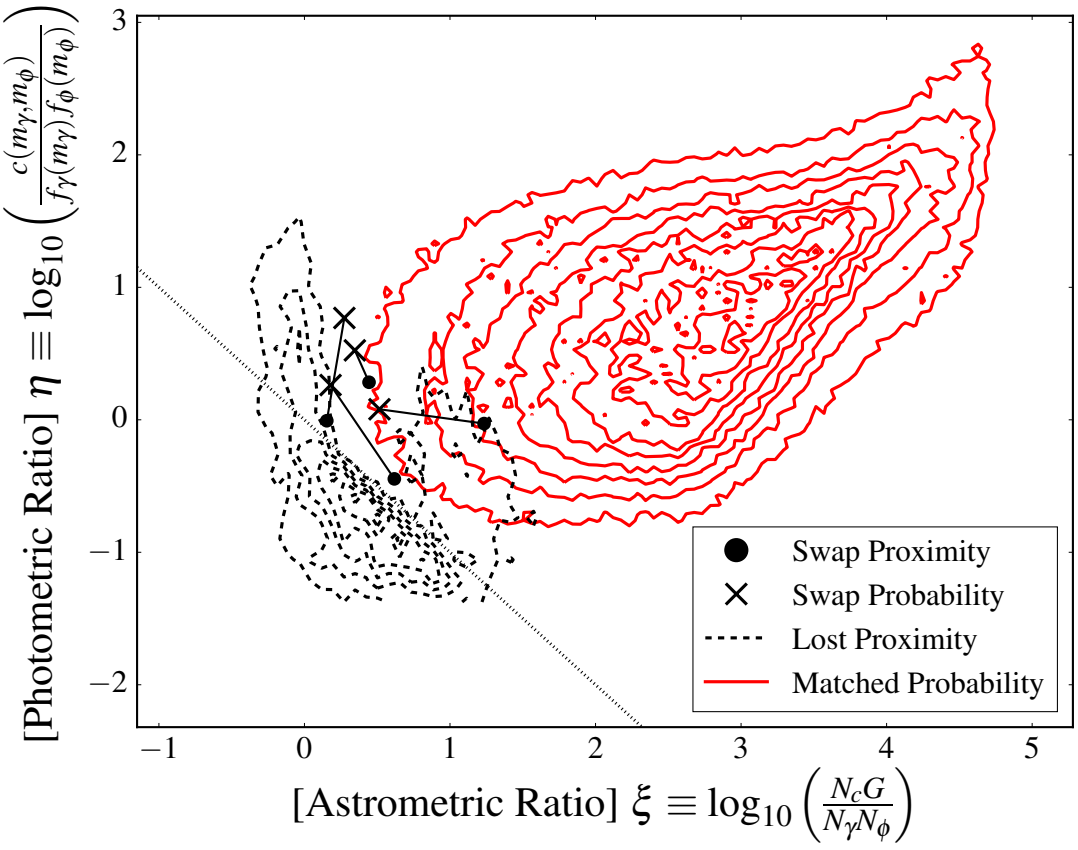


Figure 4.10: The astrometric and photometric likelihood ratios for *Gaia*-*WISE* matches for a 42 square degree region of the Galactic plane. The details of the figure are the same as Figure 4.2, with the addition of crosses and filled circles connected by a solid black line. These represent the likelihood ratios of *Gaia* objects which were nearest neighbour matched to one *WISE* object (circles) but matched to a different star through the probability-based matching process (crosses). Those objects that were nearest neighbour matched but unmatched in a probability-based match lie at slightly higher than equal chance astrometrically, but are unlikely enough photometrically to drop below a combined equal likelihood of $\xi + \eta = 0$. These matches are therefore rejected by their mismatched photometry, rather than their spatial correlation (or lack thereof). In addition, almost all of the objects which swap their returned match shown increased photometric likelihood, indicating a more likely match based on their magnitudes in the two filters.

for the use of empirical AUFs which have much larger non-Gaussian tails, on the order of several arcseconds, while simultaneously rejecting unphysical matches. Without the extra information provided by the magnitudes of the sources, the false match rate could potentially rise to unacceptable levels, resulting in untrustworthy merged datasets.

There are also some cases where stars have swapped match between the nearest neighbour- and probability-based matches. These matches increase in η, suggesting a more likely match photometrically, possibly at the expense of a small amount of astrometric likelihood. Pairings which decrease in combined likelihood ratio (i.e., $\eta + \xi$) can be explained by the fact that these ratios consider the two stars in isolation. The full matching process considers all objects in both catalogues that are spatially correlated at once. This suggests that while the new probability-based match is slightly less favourable, another match considered jointly was more favourable, overcoming the slight loss in isolated likelihood ratio.

4.4 Galactic Plane Matches

In this section I analyse the cross-matching of *Gaia* and *WISE* catalogues of photometric detections, following the probability-based matching process discussed in Chapter 3, with the addition of the construction of empirical AUFs as detailed in Section 4.3. I have chosen to focus on the Galactic plane, where the effects of crowding are most significant and thus the effects of perturbation in the *WISE* AUFs most extreme. There are a few important, but perhaps subtle, caveats that bear repeating at this point. If an object in either catalogue does not appear, not meeting the catalogue cleaning criteria, shown in Table 1.2, its counterpart will be returned unmatched. Additionally, it should be repeated that a not insignificant fraction of *Gaia* objects will not have a detected *WISE* counterpart due to their being merged inside a brighter *WISE* object's PSF. Therefore, the non-matching of a *Gaia* object should not necessarily be seen as an upper limit on the *WISE* fluxes, or vice versa.

The motivation for the creation of this formalism is, fundamentally, to provide lists

of pairs of objects between two given catalogues. To provide a good catalogue of matches, it would be useful to provide the source pairings and some key information (e.g., positions, magnitudes, names), but additional information to allow the user to evaluate whether they wish to accept the pairing. In the following discussion I follow the "Full Coverage" method outlined in Section 2.9.2.

When accepting a source pairing, here accepting the most likely match hypothesis without regard to its value in contrast with the discussion above, the probability of each of the sources being a contaminated source can also be calculated. At a given separation, the probability of a match being contaminated by an additional source of flux, denoting this hypothesis as ψ, is

$$P(\psi|r) = \frac{P(\psi)\, p(r|\psi)}{p(r)}, \tag{4.11}$$

where r is the separation between the two matched sources. For a two-directional match this equation is slightly more complex, considering the hypotheses that both objects are contaminated, one but not the other source is contaminated, and the chance that neither object is affected by systematic perturbations. Each source can be considered in turn, representing the hypotheses that a *Gaia* source is contaminated as ψ and uncontaminated as $\widetilde{\psi}$, respectively. Analagously, the hypotheses are ω and $\widetilde{\omega}$ for the cases of a contaminated and uncontaminated *WISE* source respectively. Therefore, the hypothesis that a given source is contaminated given the separation between it and its corresponding detection, denoted P_{contam} henceforth, is to be considered, both hypotheses can be marginalised over for the match in the opposing catalogue. This would give

$$\begin{aligned} P(\omega|r) &= P(\omega, \psi|r) + P\left(\omega, \widetilde{\psi}|r\right) \\ &= \frac{P(\omega)\, P(\psi)\, p(r|\omega, \psi) + P(\omega)\, P\left(\widetilde{\psi}\right) p\left(r|\omega, \widetilde{\psi}\right)}{p(r)}, \end{aligned} \tag{4.12}$$

for the hypothesis of the *WISE* source being contaminated, assuming the priors for each

catalogue suffering contamination are independent from one another. The evidence is given by the combination of all four hypotheses,

$$\begin{aligned} p(r) = p\left(r|\psi,\omega\right) P\left(\psi\right) P\left(\omega\right) + p\left(r|\psi,\widetilde{\omega}\right) P\left(\psi\right) P\left(\widetilde{\omega}\right) + \\ p\left(r|\widetilde{\psi},\omega\right) P\left(\widetilde{\psi}\right) P\left(\omega\right) + p\left(r|\widetilde{\psi},\widetilde{\omega}\right) P\left(\widetilde{\psi}\right) P\left(\widetilde{\omega}\right). \end{aligned} \tag{4.13}$$

The priors for the contamination hypotheses are simply the fraction of numerical PSF simulations (Section 4.3.1) to suffer from additional sources with a total flux ratio greater than 1% for each catalogue in turn. To evaluate each joint hypothesis' likelihood, the evaluation of the convolution of h_{pure} for both catalogues and h_{offsets} for any catalogue in which the contamination hypothesis is being considered is required. For the case above of a contaminated *WISE* source and uncontaminated *Gaia* detection, the likelihood is

$$p\left(r|\omega,\widetilde{\psi}\right) = \left(h_{\omega,\text{pure}} * h_{\psi,\text{pure}} * h_{\omega,\text{offsets}}\right)\left(r\right), \tag{4.14}$$

where the syntax $(f * g)(x)$ represents the convolution of functions f and g evaluated at x. These probabilities would aid in the selection of uncontaminated *WISE* sources. Wherever used, P_{contam} represents the probability that the source in the given catalogue in question suffers contamination above 1% relative flux given the separation between it and its corresponding detection, independent of the detection in the opposing catalogue.

4.4.1 Galactic Plane Match Testing

In Section 4.3.5 I analysed a representative region of the Galactic plane, comparing the improved empirical AUF treatment to a naive nearest neighbour match and a simplistic, pure Gaussian AUF. In this subsection I examine the matching process in more detail, discussing a variety of tests applied to the *Gaia*-*WISE* matches.

4.4.1.1 The Effect of Simulated Source Counts on Match Fractions

The first test examined is that of the effect of the simulated AUFs on the pairings obtained. Both the creation of the perturbed distribution (Section 4.3.1) and the formulation of the differential source counts used to evaluate PSF circle densities (Section 4.3.2.2) use stochastic processes, and therefore will change with each iteration. To quantify the level of variation these stochastic processes introduce, I ran two identical matches on the catalogue used in Sections 4.2 and 4.3. Of the $\approx$ 642000 matches in the region, approximately 250 pairings were not shared by both composite catalogues, on the order of 0.05% of matches. The acceptance or rejection of these matches lies in equation 4.10. Depending on the subtle variations in the empirical AUF created in each match process, these sources lie either just inside, or just outside of, the 99^{th} percentile of the AUF integral. They are therefore rejected in one run as being incompatible astrometrically, but accepted in the other. This effect is an unavoidable side effect of using empirical treatments, and should be considered carefully for cases where sources might be separated by large distances, such as high proper motion sources.

4.4.1.2 The Effect of Normalisation Radius on Match Rate

Another potential source of variation in the matching process is the local density normalisation (equation 4.9). To evaluate the level the choice of normalisation radius affected the results, I ran a match identical that used in Section 4.3.5, but with a one degree normalisation radius, rather than 15 arcminutes as is used in all other cases. I found that there were 100 matches – 0.015% of the overall matches – differing between the two matches, well within the variation due to stochastic processes used in the matching process (Section 4.4.1.1). This suggests that the density of sources does not vary in scales between one degree and 15 arcminutes, and that the evaluation of N for each source is robust. This can be seen in Figure 4.11, where the normalisation density was calculated for a variety of radii for stars in $131 \leq l \leq 138$, $-3 \leq b \leq 3$. The smaller density evaluation radii give a slightly larger spread of N compared to larger radii, but typical variations are on the

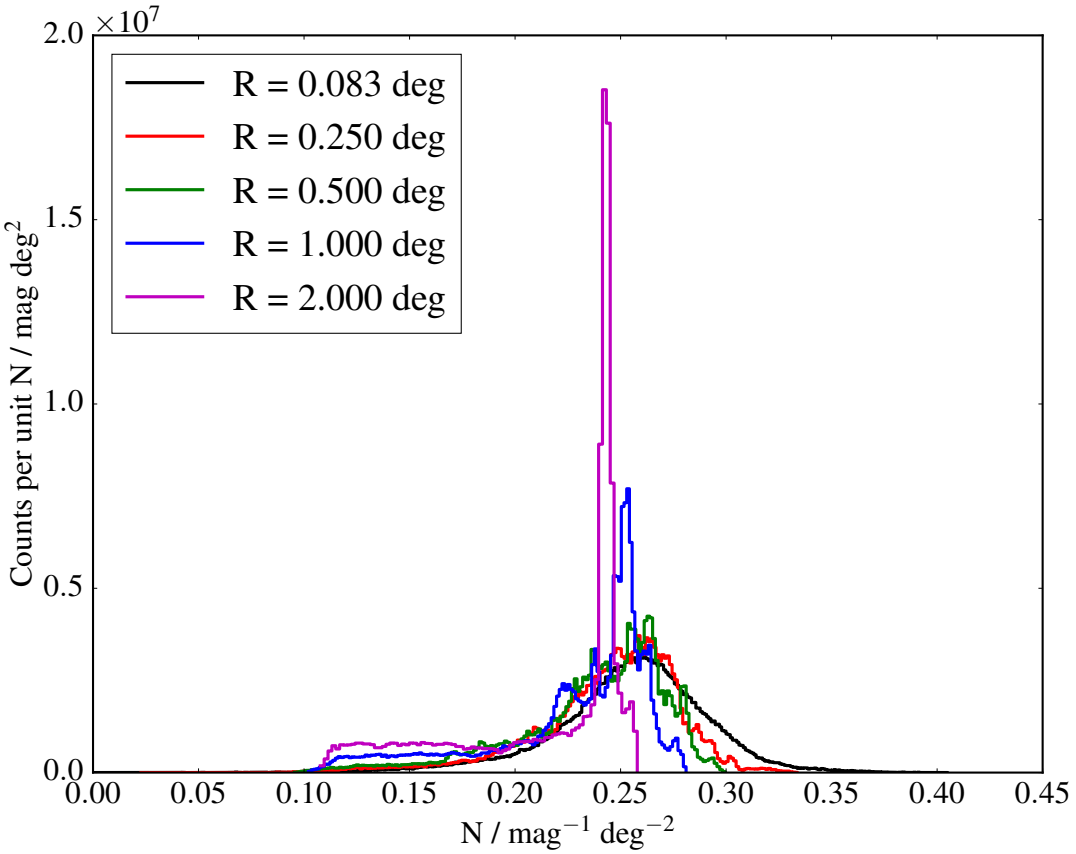


Figure 4.11: The effect of normalisation radius on the calculation of N for *WISE* stars in a 42 square degree region of the Galactic plane. The number of stars in circles of varying radii was calculated, and then N computed from equation 4.9. The effect is overall small, with larger radii simply decreasing the spread of normalisation densities, tending towards the global source density.

orders of percent. With very large radii, 1-2 degrees, the distribution of N tends towards a global average source density.

4.4.1.3 Analysis of the *Gaia-WISE* False Match Rate

A useful metric for consideration of any dataset is its false match rate. This level can be quantified by matching between a *Gaia* catalogue from one region of the Galactic plane, and a *WISE* catalogue from a second region, under the assumption that the positions of sources across the Galactic plane are independent from one another. To achieve this I took all *Gaia* sources $121 \leq l \leq 128$, $-3 \leq b \leq 3$, filtered them for quality as per the criteria in Table 1.2, then incremented their Galactic longitude by 10 degrees (i.e., if a star is recorded at $l = 125$, $b = 0$, I "moved" it to $l = 135$, $b = 0$). I then ran a match between

this new *Gaia* catalogue and the original *WISE* catalogue, returning $\approx$ 3300 matches, or 0.4% of the input *WISE* catalogue. This highlights the improvements the additional information available to the matching process, compared with a simple nearest neighbour match. A nearest neighbour match should return 4% false matches (Section 4.2.2). A factor ten improvement is seen in the false match rate, with both the variable scale length and inclusion of the photometric information allowing for the identification and rejection of 9 out of every 10 uncorrelated star pairs.

4.4.1.4 The Effect of Photometric Likelihood Inclusion on Match Fraction

The effect the inclusion of the photometric information of the catalogues has on the matches returned by the matching process can be examined further. Removing from consideration the weighting of the hypotheses by star brightnesses the pairings accepted and rejected, and the relative probabilities they are assigned, can be analysed. Setting $c(m, m) = f(m)f(m) = 1$ (see Section 3.3.2.2 for more details), the matching process returned $\approx$ 671000 matches, cf. the $\approx$ 642000 matches obtained with the photometric probability densities' inclusion of which $\approx$ 637000 matches are shared between the two matching processes. As expected, the photometric likelihood ratio being included allows for the inclusion of $\simeq$ 1% of matches, but more crucially rejects 90% of the $\simeq$ 4% of serendipitous matches expected to occur. Comparing the match probabilities for the common matches accepted by both processes the inclusion of the photometric information improves the overall probability of acceptance. Additionally, comparing the median Bayes' factor of the null hypothesis (that these two sources being unrelated detections), an increase of a factor of approximately five is found (cf. the photometric likelihood ratio η, Figure 4.10).

4.5 Discussion

4.5.1 Comparison with Literature Catalogue Matching Methods

It is useful at this point to compare the method I present here, and the results obtained, to those currently available in the literature.

4.5.1.1 Comparison with Pure Gaussian AUF Literature Matching Methods

The most obvious difference between the method laid out here, building upon the probability-based matching processes laid out in Chapter 3, and previous literature works, is the effect of relaxing of the assumption of Gaussianity in the AUF. When using a pure Gaussian AUF 55% of the sources returned with a fully empirical AUF that takes into account the effects of crowding are matched. Therefore any cross-matching method that does not take this or any additional perturbations into account will underestimate its match fraction significantly. While *WISE* is perhaps one of the more extreme cases for crowding, being a deep and complete survey with a large PSF, these effects are still non-negligible for other catalogues. For example, 2MASS (Skrutskie et al., 2006) suffers crowding at its median magnitude that causes on the order of 10% of stars to be perturbed beyond the separation where a Gaussian-only AUF would successfully recover them. Even for sources as bright as $K_s \simeq 12$ this effect is at the 3% level.

It is therefore critical that these systematic effects – perturbations due to crowding, but more generally any systematics such as proper motion, parallax, astrometric solution offsets, etc. – are included in the AUFs of these catalogues. The general formalism of the AUF derived in Chapter 3 allows for these effects to be folded in trivially; see Section 4.5.6.1 for more details. I therefore recommend the reader consider the catalogues being cross-matched, particularly with reference to the typical density, sources per PSF circle, before accepting the results of any cross-match involving a pure Gaussian AUF (e.g., Sutherland and Saunders, 1992 and any work building upon their "LR" method, such as Pineau et al., 2017; Budavári and Szalay, 2008; Salvato et al., 2018; or Marrese et al.,

2017).

4.5.1.2 Direct Match Comparison with *Gaia* DR1

A more direct comparison to a literature cross-match can be performed, comparing the matches I obtain to those provided as part of the *Gaia* DR1 release (Marrese et al., 2017). As part of the release they provide cross-matches between *Gaia* and *WISE*, allowing for an analysis of the matches between the two methods. My method returns 82% of all *WISE* sources as being matched to a *Gaia* source in the 42 square degrees of the Galactic plane centered on $l = 135$, $b = 0$, in good agreement with the official *Gaia* DR1 match fraction (figure 3n, Marrese et al., 2017). However, the extra matches they obtain, compared with the match rate I find in Section 4.2, are a result of the broadening of their astrometric uncertainties (section 3.2 of Marrese et al., 2017), which they believed accounted for epoch differences and any resultant proper motion shifts of the sources. These broadened astrometric uncertainties are much larger than the typical precision of either dataset, leading to the case where the parameters of the Gaussian AUF are independent of the properties of the sources themselves. The approximately constant uncertainties lead, effectively, to a reduction to a nearest neighbour match, with a matching radius that depends on the local source density. This radius, in most cases, is sufficiently large to capture the non-Gaussian wings of the full AUF, resulting in most pairings successfully being recovered. Their analytical solution is useful, allowing for simpler computations and the flexibility to include the relative likelihood of multiple matches. Marrese et al. (2017) use this advantage to assign multiple *Gaia* "mates" to singular *WISE* counterparts, accounting for the higher *Gaia* angular resolution deblending otherwise confused sources. However, the uncertainty broadening required to provide a good match rate, overcoming the astrometric perturbation from this crowding, has another, more subtle effect.

The effect in question can be explained as follows. The astrometric uncertainty broadening in turn reduces the maximum probability density of the Gaussian, being a normalised function, which has implications for null hypothesis testing. To test this

I obtained the *Neighbourhood* results for the *Gaia-WISE* matches from Marrese et al. (2017). I converted these scores to "Figures of Merit" (FoM), multiplying the figures of merit by a factor 3600 (P. Marrese, priv. comm.). In six cases there were multiple *mates* for *Gaia* sources, for which I picked the largest FoM (see Marrese et al., 2017 for details). The "reliability" (Sutherland and Saunders, 1992), or the normalised probability of the pairing hypothesis, including the null hypothesis (or the two sources being uncorrelated and detections of differing objects) was then obtained by

$$P(r) = (1 + (FoM(r))^{-1})^{-1}. \tag{4.15}$$

The probability obtained using the method laid out in Section 4.3 was then compared to those given as part of the *Gaia* DR1 release. 85% of sources in each individual catalogue are shared between both – likely caused by differing quality cuts between the two *Gaia* catalogues used – and their probabilities are compared in Figure 4.12. As can be seen, the broadening of the astrometric uncertainties leads to the most certain matches having a constant, but much lower, probability for the Marrese et al. (2017) matches, compared with those presented in this chapter. This constant but reduced Gaussian probability density results in a reduction in the confidence with which the non-match hypothesis can be rejected, by up to five orders of magnitude in some cases.

4.5.1.3 Perturbation Offset Determination Comparison

In this chapter I deal with the effects of contaminant star perturbation by calculating flux-weighted centroid shifts to the central source, following the method discussed in Chapter 2. The applicability of these centroid shifts depends on the data reduction scheme applied to the images of the given observations. The flux-weighted centroid scheme is appropriate when positions have been found by centroiding, usually followed by aperture photometry to calculate the flux of the detection. However, there are data reduction schemes where PSF fitting is undertaken to calculate source fluxes and positions, the main method utilised to reconstruct sources in the *WISE* data releases. In this instance the difference between

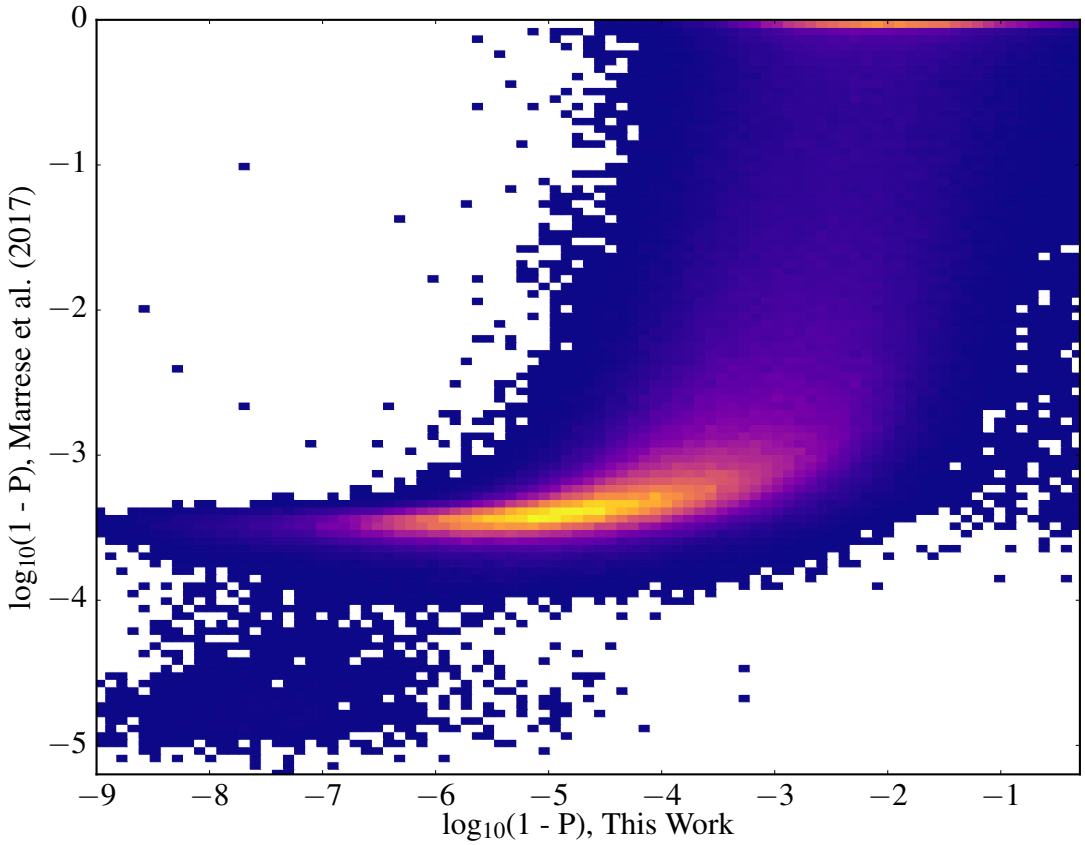


Figure 4.12: Comparison between match probabilities of *Gaia* sources, as calculated using the method outlined in Section 4.3 and the probability calculated from the "Figures of Merit" as quoted by Marrese et al. (2017). The colour scale shows the number of stars in each two-dimensional bin, with blue representing low counts and yellow high counts. The astrometric broadening of the Gaussian uncertainties used by Marrese et al. (2017) lead to a plateauing of the probabilities for the most certain matches, with as large as five orders of magnitude difference between the confidence with which the two methods reject the null hypothesis. The increase in probability for the small group of less certain matches is likely caused by the "island" creation process discussed in Section 3.6.2, leading to an increase in certainty of hypotheses when multiple pairs are simultaneously fit between the two catalogues.

two PSFs – the bright source and the faint contaminating source – and a slightly brighter, slightly shifted PSF representing the blended object should be minimised when evaluating potential perturbations. Plewa and Sari (2018) use this method to explore the effects of confusion on the orbits of S-stars in the central few square arcseconds at the Galactic centre.

However, as Plewa and Sari (2018) show, the analytical approximation to this minimisation only gives good agreement to the full solution for $\Delta m \geq 3$, or flux ratios less than approximately 6%. Using the full numerical solution is computationally intractable for large-scale catalogue matching, and thus if this alternate method is considered the analytical expression would have to be used. For stars sufficiently bright that this inequality is valid, with typical contaminating sources at least three magnitudes fainter than the central source, I found that the centroiding and PSF fitting methods produce empirical AUFs that are in reasonable agreement with one another. I tested the offset perturbations produced by both methods against *Gaia-WISE* separation distributions for *WISE* stars $131 \leq l \leq 138$, $-3 \leq b \leq 3$, $W1 = 13.47 - 13.53$, $\sigma_\alpha = 0.023 - 0.083$ arcsecond, and $N = 0.257 - 0.267\,\mathrm{mag}^{-1}\,\mathrm{deg}^{-2}$. Both methods produced AUFs which fit the non-Gaussian tails to the separations, with fits to the full cross-match separations of $\chi^2_\nu \simeq 1.7$ for the PSF fitting method and $\chi^2_\nu \simeq 2.4$ for the flux-weighted centroid method, respectively, with zero free parameters.

WISE suffers extreme levels of crowding, however, and is potentially flux contaminated on the order of 15% for stars as bright as $W1 \simeq 12$ (see Section 4.5.2.2). Thus in regions of extreme crowding, or catalogues that are especially affected by crowding, such as *WISE*, the flux-weighted centroid method produces AUFs much closer to the distribution of source separations than the analytical expression to the PSF fitting method derived by Plewa and Sari (2018). I tested this using the same sky region and normalisation density cut as before, but with *WISE* stars in the range $W1 = 15.47 - 15.53$ and $\sigma_\alpha = 0.093 - 0.153$ arcsecond. I found the flux-weighted centroid offset calculations produce offset distributions that result in an AUF with $\chi^2_\nu \simeq 2.4$ when compared with the cross-match separation

distribution, as previously, but the analytical approximation to the PSF fitting resulted in a much larger goodness-of-fit, $\chi^2_\nu \simeq 6.3$. Therefore, while the flux-weighted centroiding method does not reflect the data reduction process as closely as the PSF fitting method, it produces AUFs in good agreement to the separations seen across all magnitudes. The analytical approximation to PSF fitting description does not hold for the majority of the *WISE* stars, however. I therefore use the flux-weighted centroid method for the creation of the *Gaia*-*WISE* matches I discuss here.

4.5.2 Photometry Differences

So far I have discussed the effects faint, hidden stars have on the astrometric positions of sources. Simultaneously, they also introduce additional flux to the central source. In this section I will discuss the effect crowding has on the photometry of blended sources, showing that the correct treatment of the astrometry of sources can reveal the introduction of additional brightness into these perturbed sources.

4.5.2.1 The Effect of Perturbation on *WISE* Brightnesses

The first test that can be done to examine the effect crowding has on the flux contamination is to compare the two matching cases. In effect, there are two distributions in the dataset of counterparts returned by the empirical AUF matching process used in Section 4.3. First, those objects whose astrometric separations in *WISE* and *Gaia* are compatible with a Gaussian AUF, and would therefore have been matched in the Gaussian-based match in Section 4.2. Second, the subset of objects with AUFs incompatible with a Gaussian, perturbed to the level that they were rejected by the Gaussian AUF matching process. These objects must have a hidden contaminant of sufficient brightness offset from the central source by a large enough radius such that the flux-weighted average position is beyond that allowed by the Gaussian AUF. In this subsection I will contrast these two subsets, to examine the effect this high level of perturbation has on the measured *WISE* fluxes.

Comparing the two distributions the immediate difference is the number of returned matches. The Gaussian-based match returns 55% of the matches returned by the empirical AUF. However, more significant is the statistical relationship between the G magnitude and *WISE* bands (e.g., W1). Figure 4.13 shows the distribution of $G-W1$ colours for those objects recovered with a Gaussian-based match (dashed lines), and the additional objects that are paired when the empirical AUF is employed (solid lines), for several slices in *Gaia* magnitude. With increasing G magnitude, the $G-W1$ colour shift between the matches obtained with the purely Gaussian AUF and the additional empirical AUF-only matches increases. The objects gained when using an AUF that includes large, non-Gaussian wings are on average 0.26 magnitudes brighter in W1 for the same *Gaia* magnitude than those recovered with a Gaussian AUF. This implies that the average flux contamination leading to these large wings, in those objects not captured by the Gaussian AUF due to sufficient flux contamination, is approximately 27%, similar to the average flux contamination of 23% seen in the empirical AUFs created in Section 4.3.

To test further whether there was a correlation between photometric contamination and astrometric perturbation I divided the set of additional matches into two subsets, split by median sky separation. Fitting the $G-W1$ relationship of both halves of the gained matches I found a trend with $W1$ – or, equivalently, G – magnitude, shown in Figure 4.14. Following the same G magnitude slices as Figure 4.13, now plotted are matches separated by less than (solid lines) or more than (dashed lines) the median separation. At faint magnitudes ($W1 \simeq 16$, or $G \simeq 19$) there is an inverse trend with match separation, with objects at smaller match separations exhibiting systematically more flux contamination. However, at increasingly bright magnitudes ($W1 \lesssim 13$, or $G \lesssim 15$) the objects with high astrometric perturbation are on average more flux contaminated than those objects that do not show high perturbation. At faint magnitudes, and thus high stellar densities, there are multiple contaminants in each *WISE* PSF. As additional flux contamination from increasing numbers of contaminant stars is added, the overall flux-weighted centroid will tend towards zero. Therefore, in this high effective density regime the highest perturbations

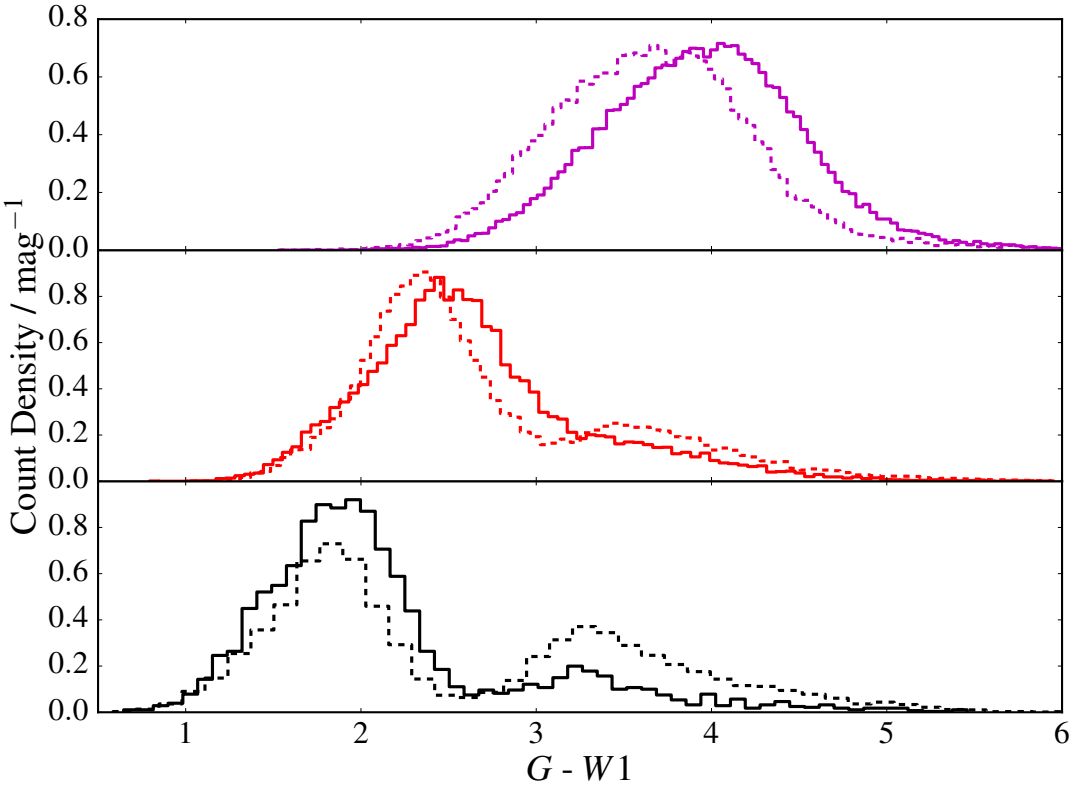


Figure 4.13: The $G-W1$ colour of *Gaia-WISE* matches that were paired using Gaussian-based AUFs (dashed lines) and the additional *Gaia-WISE* matches recovered using an empirical AUF (solid lines). Shown are the matches for stars with $12 \leq G \leq 14$ in black (bottom panel), $15 \leq G \leq 16$ in red (middle panel), and $19 \leq G \leq 20$ in magenta (top panel). The shift in $G - W1$ colour for those additional, empirical-only matches increases with increasing G magnitude, suggesting an increasing $W1$ contamination. The average W1 magnitude is 0.26 magnitudes brighter for the non-Gaussian matches compared to those that are matched with a Gaussian AUF, implying $\simeq 27\%$ flux contamination, comparable to the average contamination seen in the constructed empirical AUFs.

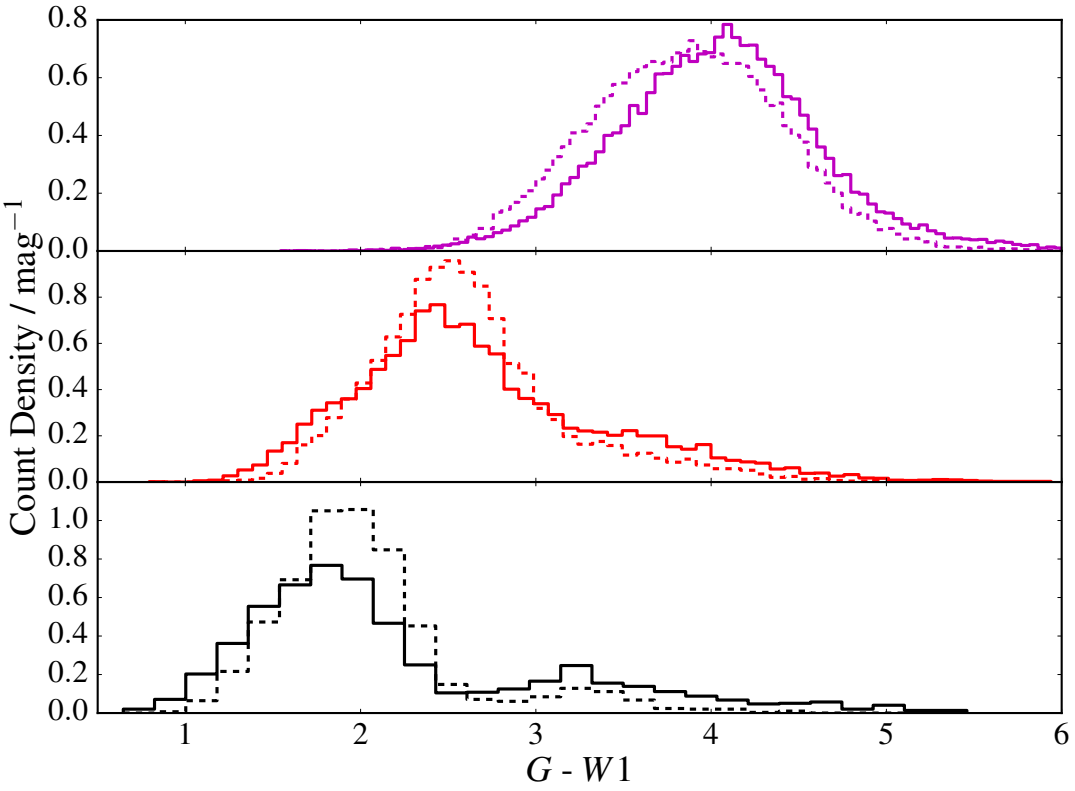


Figure 4.14: The $G - W1$ colour of the additional *Gaia-WISE* matches recovered using an empirical AUF as a function of sky separation. The colours of those matches recovered with sky separations below (solid lines) and above (dashed lines) the median separation for the matches in the magnitude range are plotted. Shown are the matches for stars with $12 \leq G \leq 14$ in black (bottom panel), $15 \leq G \leq 16$ in red (middle panel), and $19 \leq G \leq 20$ in magenta (top panel). At bright magnitudes the contamination increases with increasing sky separation, suggesting a trend with contaminant distance. At faint magnitudes the effect is reversed, with multiple contaminants resulting in a flux-weighted centroid closer to zero with increasing number of contaminants and thus flux contamination.

are seen in sources with lower levels of flux contamination, caused by a smaller number of faint sources. However, for brighter objects the effective stellar density is reduced, which leads to on average one contaminant that can affect the recorded position. This then results in a situation where there is a correlation between measured offset and contaminant brightness, as observed.

4.5.2.2 Resolving Contaminants with *Spitzer*

To confirm whether any source is contaminated, the matches in a higher angular resolution dataset can be examined. *WISE*'s $W1$ and $W2$ bands have very similar coverage to *Spitzer*'s IRAC (Fazio et al., 2004) 3.6μm and 4.5μm bands, offering a resolution of

≃2 arcsecond FWHM. I therefore obtained *Spitzer* Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) data in the region $131 \leq l \leq 138$, $0 \leq b \leq 2$, and constructed empirical AUFs (see Section 4.3.2) for the two IRAC filters available. I then performed a probability-based matching procedure to *Gaia*, as detailed in Chapter 3. The assumption was made that stars in both mid-infrared datasets that matched to the same *Gaia* object were the same source detected at two different epochs in the two catalogues. I selected stars $11.5 \leq W1 \leq 12$, bright enough that *WISE* sources are not entirely dominated by contamination, allowing for comparison between contaminated and uncontaminated sources.

Two subsets of these common *Gaia* matches were obtained: likely uncontaminated *WISE* objects and likely contaminated *WISE* objects, based solely on the *Gaia*-*WISE* matching. These correspond to $P_{\text{contam}} \leq 0.25$ and $P_{\text{contam}} \geq 0.85$ (equation 4.12) respectively. Once these four subsets (*WISE* and *Spitzer* objects which correspond to both contaminated and uncontaminated *WISE* objects) were obtained, the intra-catalogue separation (i.e., the distance to the nearest *WISE* object for a given subset of *WISE* objects) was found. I limited the intra-catalogue search to stars with brightnesses $m \leq 15$ in both catalogues, allowing for consistent testing. Without the magnitude limit *Spitzer*'s fainter completeness limit would otherwise have resulted in a smaller average offset than for that of *WISE*, caused by an increase in the number of stars in any given region. The distribution of intra-catalogue separations for *Spitzer* is shown in Figure 4.15.

For the *Spitzer* objects corresponding to uncontaminated *WISE* objects (dashed line), the distribution of separations to the nearest intra-catalogue object (i.e., the nearest other *Spitzer* detection) corresponds to the typical distance between sources at the given stellar density, approximately 25 arcseconds. This gives good agreement with both the contaminated and uncontaminated *WISE* intra-catalogue distributions. However, the *Spitzer* objects which correspond to contaminated *WISE* objects (solid line) show a different distribution. With the better angular resolution *Spitzer* has the ability to resolve two objects previously blended in *WISE*. The nearest *Spitzer* neighbour is therefore likely to

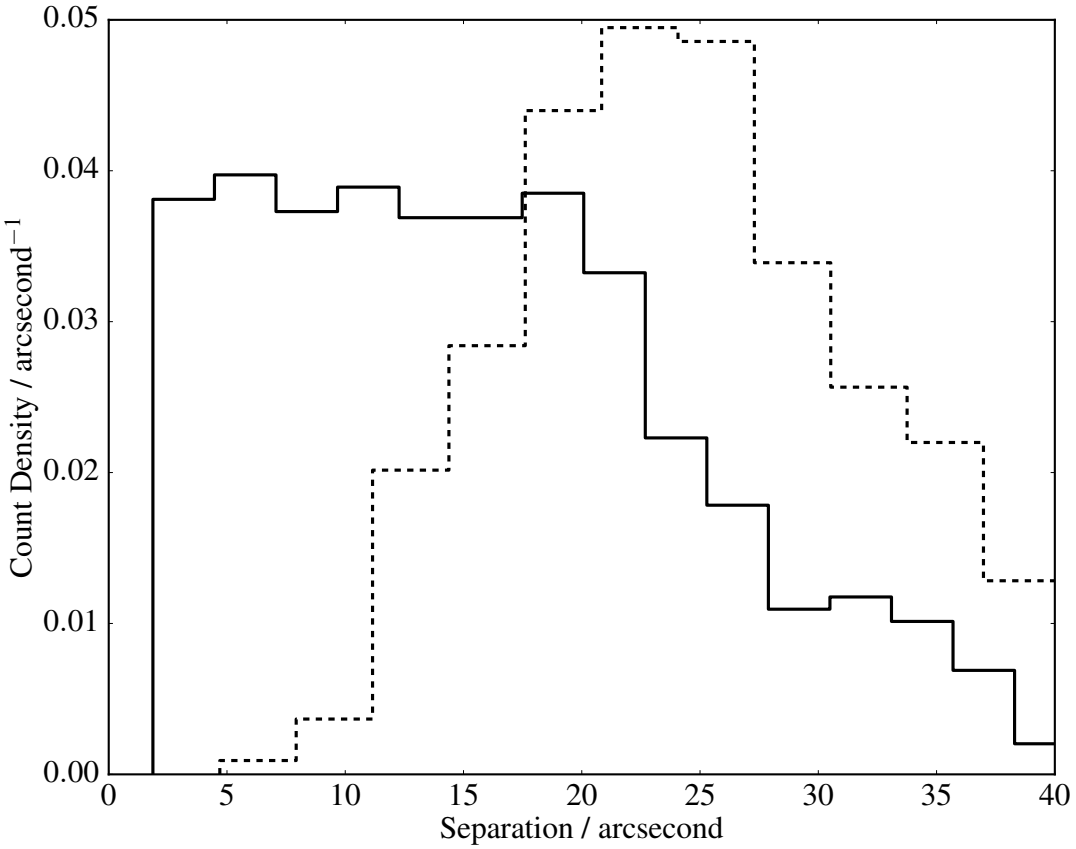


Figure 4.15: The intra-catalogue nearest neighbour distances for two samples of *Spitzer* stars. Shown are those stars with a common *Gaia* source to *WISE* sources $11.5 \leq W1 \leq 12$. The two cases are those matches where the *WISE* matches are unlikely to be contaminated ($P_{\rm contam} \leq 0.25$; dashed lines), and the case where the *WISE* objects have a high probability of contamination ($P_{\rm contam} \geq 0.85$; solid lines). The faintest magnitude for intra-catalogue separation consideration was limited to 15, to account for *Spitzer*'s fainter completeness limit. The *Spitzer* detections of uncontaminated *WISE* objects share a similar nearest neighbour distance distribution with both contaminated and uncontaminated *WISE* sources. However, the *Spitzer* nearest neighbour distribution for contaminated *WISE* objects shows a much smaller average offset, with *Spitzer* resolving the *WISE* contaminants.

be the hidden *WISE* contaminant, as shown by a distribution skewed towards separations $\lesssim$ 10 arcsecond.

This resolving of contaminants is further confirmed when the magnitude differences between the *WISE* and *Spitzer* objects are compared for the two sources, similar to Section 4.5.2.1. For the uncontaminated *WISE* objects, the median $W1 - [3.6]$ colour is 0.005 magnitudes, while the subset of sources with significant *WISE* contamination have a median $W1 - [3.6]$ of -0.132 magnitudes. This implies that, even as bright as $W1 \simeq 12$, some *WISE* sources are suffering flux contamination on the order of 15%.

4.5.3 The Effects of Invisible Perturbants

While there is good agreement between the empirical AUF constructed following the method laid out in Sections 4.3.1 and 4.3.2 and the distribution of separations between sources in the two catalogues, the effect of not including a more detailed treatment can be highlighted here. The red dashed line in Figure 4.7 shows the empirical AUF obtained if the full treatment of the differential source counts is not taken into account (i.e., Nz^m is assumed to continue to arbitrarily faint magnitudes). As can be seen, this AUF does not fit the distribution of separations correctly; however, the magnitude of the central sources is almost at the sensitivity limit of the survey.

This means that the vast majority of sources affecting the AUF and the perturbation of the bright sources would not be detected by the survey in a sparse field. This highlights the importance of the correct treatment of the density of faint contaminants. If treated correctly, the effects of otherwise "invisible" stars can be seen indirectly in their influence on brighter objects.

4.5.4 Circular Symmetry in Empirical AUF Creation

The formalism given here for the creation of empirical AUFs implicitly assumes circular symmetry. I have assumed a circular PSF in the previous sections, and for the discussion in Sections 4.3.5 and 4.5.2 I additionally assume the astrometric uncertainties are circular

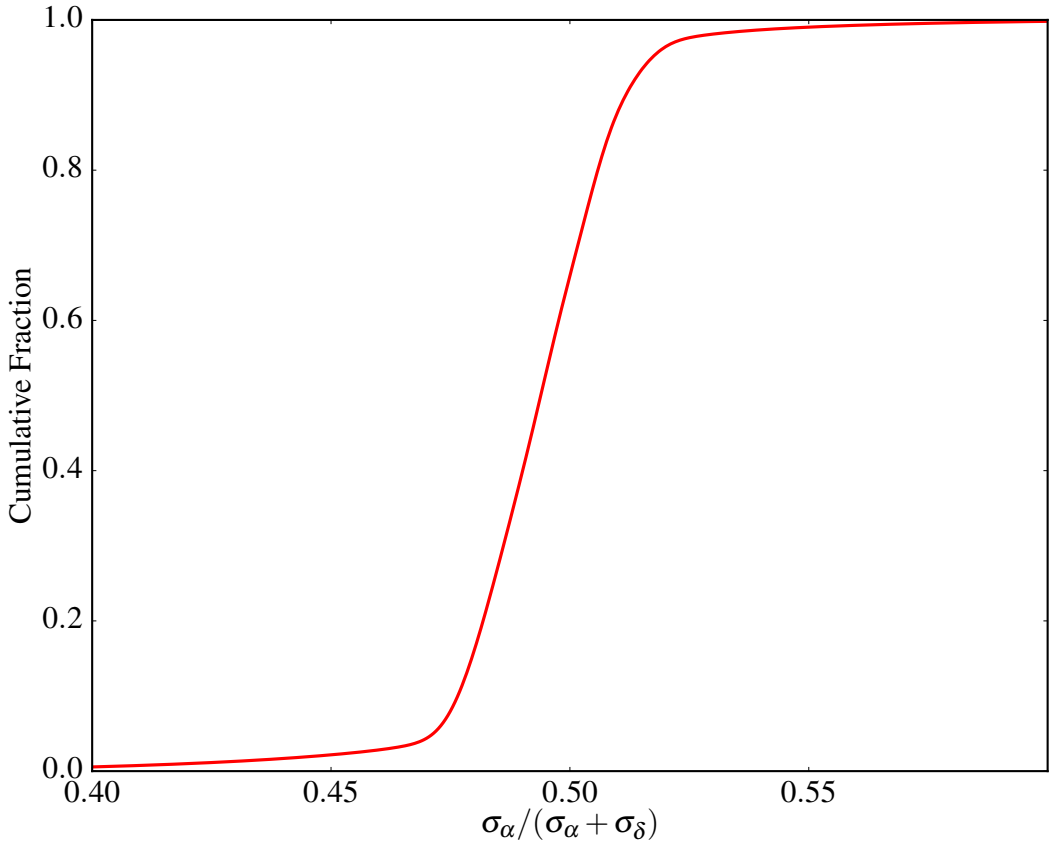


Figure 4.16: The cumulative counts of the ratio of orthogonal sky axis uncertainties, for *WISE* stars with $|b|\leq 10$. A circular positional uncertainty has a ratio of 0.5.

(i.e., $\rho = 0$). This assumption holds for the majority of sources, as ground-based surveys should have circular PSFs and thus circular centroiding uncertainties. Space-based observations, such as those for *WISE*, can have off-axis correlations in their PSFs, and thus position uncertainties, however. In practice, this effect is limited and 90% of the *WISE* data discussed here have orthogonal sky axis uncertainties that deviate from circular by less than 10%, as shown in Figure 4.16. If the ratio is to be more than 10% different, then $\sigma_\alpha/(\sigma_\alpha + \sigma_\delta)$ would have to be smaller than 0.474 or larger than 0.524. Therefore, while the convolution of the distribution of perturbations and a Gaussian preserving the full covariance matrix is possible, the loss of information is negligible, vastly outweighed by the simplifications the assumption allows.

The advantage of this is that either equatorial – α, δ – or Galactic – l, b – coordinates can be used with the assumption of circularity of the positions. It can be proven that the

uncertainty in the position is constant for both coordinate systems. Using the notation defined by Lindegren et al. (2016) and van Altena (2012), the Cartesian representations of the two coordinate systems are

$$\mathbf{r}_{\rm eq} = \begin{bmatrix} X_{\rm eq} \\ Y_{\rm eq} \\ Z_{\rm eq} \end{bmatrix} = \begin{bmatrix} \cos(\alpha)\cos(\delta) \\ \sin(\alpha)\cos(\delta) \\ \sin(\delta) \end{bmatrix} \quad \mathbf{r}_{\rm Gal} = \begin{bmatrix} X_{\rm Gal} \\ Y_{\rm Gal} \\ Z_{\rm Gal} \end{bmatrix} = \begin{bmatrix} \cos(l)\cos(b) \\ \sin(l)\cos(b) \\ \sin(b) \end{bmatrix}. \tag{4.16}$$

The transformation is defined through the relationship

$$\mathbf{r}_{\rm Gal} = \mathbf{A}'_{\rm G}\mathbf{r}_{\rm eq}, \tag{4.17}$$

with the matrix $\mathbf{A}'_{\rm G}$, the transpose of the matrix $\mathbf{A}_{\rm G}$ formalised for the Hipparcos catalogue (ESA, 1997), defined as

$$\begin{aligned} \mathbf{A}'_{\rm G} &= \begin{bmatrix} A_1 A_2 A_3 \\ A_4 A_5 A_6 \\ A_7 A_8 A_9 \end{bmatrix} \\ &= \begin{bmatrix} -0.0548755604162 & -0.8734370902349 & -0.4838350155487 \\ +0.4941094278756 & -0.4448296299600 & +0.7469822444972 \\ -0.8676661490190 & -0.1980763734312 & +0.4559837761750 \end{bmatrix}. \end{aligned} \tag{4.18}$$

The Galactic coordinates are then recovered from the transformed Cartesian representation by

$$l = \mathrm{atan2}(Y_{\rm Gal}, X_{\rm Gal}), \quad b = \mathrm{atan2}\left(Z_{\rm Gal}, \sqrt{X^2_{\rm Gal} + Y^2_{\rm Gal}}\right), \tag{4.19}$$

and thus the equatorial coordinates are converted into Galactic coordinates.

The propagation of the uncertainties from the equatorial to Galactic coordinates

requires the equatorial covariance matrix,

$$\mathbf{C}_{\rm eq} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_\alpha\sigma_\delta\rho_{\alpha\delta} \\ \sigma_\alpha\sigma_\delta\rho_{\alpha\delta} & \sigma_\delta^2 \end{bmatrix}. \tag{4.20}$$

The covariance matrix of the Galactic coordinate frame can be obtained using the transformation

$$\mathbf{C}_{\rm Gal} = \mathbf{J}\mathbf{C}_{\rm eq}\mathbf{J}'. \tag{4.21}$$

For the case in question, the preservation of the circular covariance matrix, the equatorial covariance matrix can be simplified as $\mathbf{C}_{\rm eq} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Thus

$$\mathbf{C}_{\rm Gal} = \sigma^2\mathbf{J}\mathbf{I}\mathbf{J}' = \sigma^2\mathbf{J}\mathbf{J}'. \tag{4.22}$$

The criterion for self-similarity between the equatorial and Galactic covariance matrices in the case of circularity is therefore that $\mathbf{J}\mathbf{J}' = \mathbf{I}$, or that the Jacobian is orthogonal.

$\mathbf{J}$, the transformation Jacobian, is expanded as

$$\mathbf{J} = \frac{\partial(l,b)}{\partial(\alpha,\delta)} = \begin{bmatrix} \frac{\partial l}{\partial\alpha} & \frac{\partial l}{\partial\delta} \\ \frac{\partial b}{\partial\alpha} & \frac{\partial b}{\partial\delta} \end{bmatrix} = \begin{bmatrix} \mathbf{p}'_{\rm Gal} \\ \mathbf{q}'_{\rm Gal} \end{bmatrix} \mathbf{A}'_{\rm G} \begin{bmatrix} \mathbf{p}_{\rm eq} & \mathbf{q}_{\rm eq} \end{bmatrix}. \tag{4.23}$$

with

$$\mathbf{p}_{\rm eq} = \begin{bmatrix} -\sin(\alpha) \\ \cos(\alpha) \\ 0 \end{bmatrix}, \quad \mathbf{q}_{\rm eq} = \begin{bmatrix} -\cos(\alpha)\sin(\delta) \\ -\sin(\alpha)\sin(\delta) \\ \cos(\delta) \end{bmatrix}, \tag{4.24}$$

and

$$\mathbf{p}_{\rm Gal} = \begin{bmatrix} -\sin(l) \\ \cos(l) \\ 0 \end{bmatrix}, \quad \mathbf{q}_{\rm Gal} = \begin{bmatrix} -\cos(l)\sin(b) \\ -\sin(l)\sin(b) \\ \cos(b) \end{bmatrix}. \tag{4.25}$$

Here $\mathbf{p}$ and $\mathbf{q}$ represent vectors pointing in the direction of increasing coordinates, represented in the Cartesian reference frame. The orthogonality proof can be simplified by considering that the multiplication of N orthogonal matrices will produce an orthogonal matrix. $\mathbf{A}_{\rm G}$, and therefore $\mathbf{A}'_{\rm G}$, is orthogonal, with $A_1^2 + A_2^2 + A_3^2 = 1$, $A_4^2 + A_5^2 + A_6^2 = 1$, $A_7^2 + A_8^2 + A_9^2 = 1$, and all off-diagonal sums cancelling. The orthogonality of the column vectors $\mathbf{p}$ and $\mathbf{q}$ can be proven, as

$$\begin{bmatrix} \mathbf{p}'_{\rm eq} \\ \mathbf{q}'_{\rm eq} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\rm eq} & \mathbf{q}_{\rm eq} \end{bmatrix} = \begin{bmatrix} -\sin(\alpha) & \cos(\alpha) & 0 \\ -\cos(\alpha)\sin(\delta) & -\sin(\alpha)\sin(\delta) & \cos(\delta) \end{bmatrix} \begin{bmatrix} -\sin(\alpha) & -\cos(\alpha)\sin(\delta) \\ \cos(\alpha) & -\sin(\alpha)\sin(\delta) \\ 0 & \cos(\delta) \end{bmatrix}. \tag{4.26}$$

This can be expanded as

$$\begin{aligned} &\begin{bmatrix} \mathbf{p}'_{\rm eq} \\ \mathbf{q}'_{\rm eq} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{\rm eq} & \mathbf{q}_{\rm eq} \end{bmatrix} \\ &= \begin{bmatrix} \sin^2(\alpha) + \cos^2(\alpha) & \sin(\alpha)\cos(\alpha)\sin(\delta) - \sin(\alpha)\cos(\alpha)\sin(\delta) \\ \sin(\alpha)\cos(\alpha)\sin(\delta) - \sin(\alpha)\cos(\alpha)\sin(\delta) & \cos^2(\alpha)\sin^2(\delta) + \sin^2(\alpha)\sin^2(\delta) + \cos^2(\delta) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \sin^2(\delta) + \cos^2(\delta) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned} \tag{4.27}$$

and thus $\mathbf{p}_{\rm eq}$ and $\mathbf{q}_{\rm eq}$ are orthogonal, with analagous arguments for $\mathbf{p}_{\rm Gal}$ and $\mathbf{q}_{\rm Gal}$; therefore $\mathbf{J}$ is orthogonal, and $\mathbf{C}_{\rm Gal} = \mathbf{C}_{\rm eq}$. The original circular astrometric uncertainty is therefore preserved with a change from equatorial to Galactic coordinates.

4.5.5 Extreme Crowding

The method I have outlined in this chapter accounts for the blending of sources, including the effects the brightening of the brightest source has on its astrometry. However, one of

the assumptions made was that the local density of each source could be calculated from a consistent geometric scaling relationship. As shown in Figure 4.17, this assumption may not necessarily hold in regions of extreme crowding. The black solid line shows *WISE* differential source counts of a 4 square degree region of the inner Galactic centre ($l = 0, b = 0$). Compared with the blue dashed and red dotted lines, representing differential source counts at $l = 355$, $b = 5$ and $l = 135$, $b = 0$ respectively, the Galactic centre suffers such extreme flux contamination that its brightest sources no longer follow a geometric scaling relationship. In cases where the density of sources, i.e., number of stars per PSF, is extreme, I recommend analysing the differential source counts for the catalogue in question to ensure the assumptions made about the scaling law relationship are still valid.

4.5.6 Extensions to the AUF

4.5.6.1 Extending the Empirical AUF to Additional Systematic Perturbations

In this chapter I have chosen to only include the systematic effects of crowding in my AUF treatment, being the most dominant source of non-Gaussianity in the *WISE* AUFs (see Chapter 2). However, AUFs can include any source of systematic perturbation without loss of generality. Other effects such as proper motion can be included, described by

$$h_{\rm tot} = h_{\rm pure} * h_{\rm offsets} * h_{\rm pm}. \tag{4.28}$$

Here $h_{\rm pm}$ is a probability density function describing the statistical distribution of proper motions for the catalogue in question. This has the potential to model the effects of proper motion on a large scale, in cases where individual measurements are unavailable. For example, stars fainter than *Gaia* in the next generation of photometric surveys, such as LSST, will likely lack robust individual proper motion measurements. Modelling their effects will therefore rely on such large scale statistical proper motion simulations.

The ability to include the distribution of the proper motions of sources, rather than

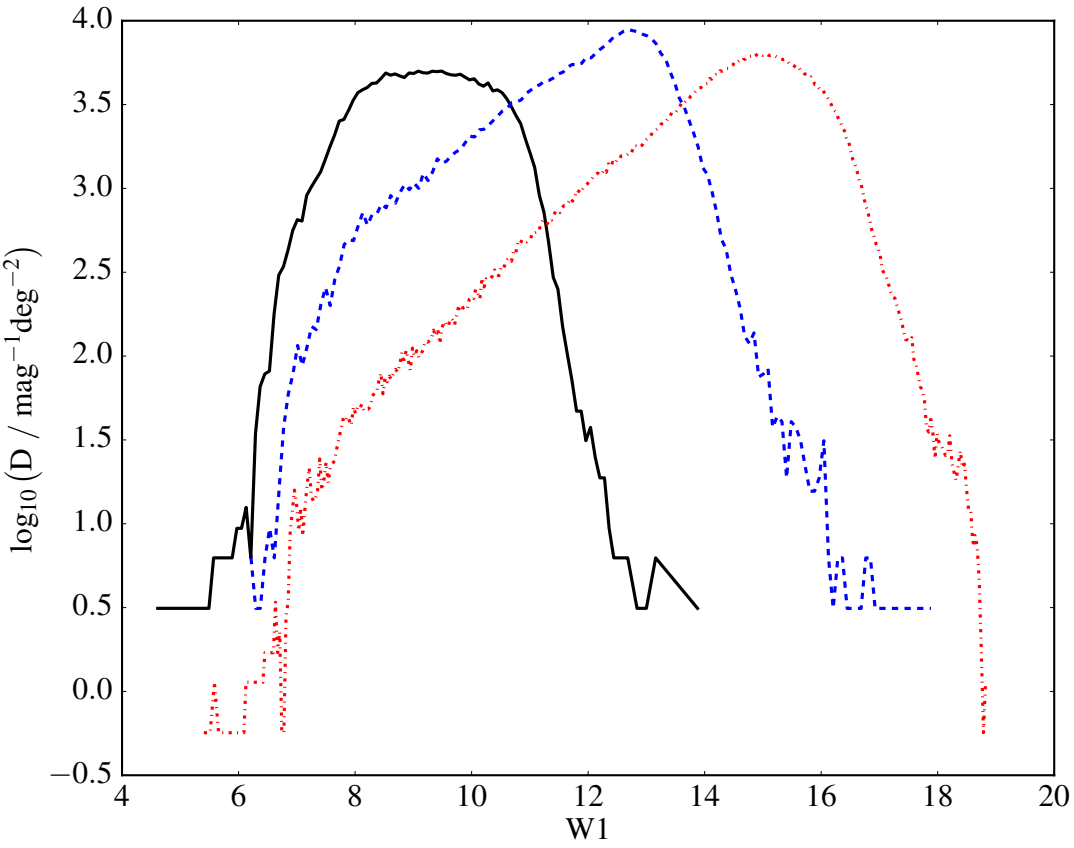


Figure 4.17: *WISE* differential source counts. The red dotted line is the differential source counts for the region of the Galactic plane around $l = 135$, $b = 0$ discussed in Section 4.3.5. The blue dashed and black solid lines are the differential source counts of 4 square degree regions of the Galactic plane at $355 \leq l \leq 357$, $5 \leq b \leq 7$ and $0 \leq l \leq 2$, $0 \leq b \leq 2$ respectively. The blue dashed and red dotted lines show relationships that follow a $z = 2$ scaling law. However, the crowding in the inner region of the Galactic centre is so extreme that the brightest sources are flux contaminated to such an extent that they no longer follow a geometric scaling law.

merely inflating the astrometric uncertainty of the position centroiding (e.g., Marrese et al., 2017), allows for a more realistic treatment of these systematic perturbations to source positions. It can be extended to be a function of multiple parameters of the catalogue – the primary one being brightness, with fainter objects having smaller proper motions on average – and does not erase the knowledge of the original positional precision. There are several cases in the literature where the motion of sources is included in the cross-matching of catalogues. Pineau et al. (2017) include an appendix discussing extending their maximum-likelihood Gaussian AUF treatment to the inclusion of the motion of sources between catalogue epochs. Similarly Kerekes et al. (2010) extend the Gaussian AUF Bayes factor method of Budavári and Szalay (2008) to account for unknown proper motions, including a more detailed treatment of the likely astrophysical proper motions of the sources as a prior term. I believe the inclusion of the proper motion offset term as part of the likelihood, inside the combined AUF, to be a more intuitive interpretation to the positional offset between catalogue source detections. It simply continues the extension to non-Gaussian perturbations, adding all terms required to correctly interpret the separations between counterpart detections to astrophysical sources.

If motions for individual sources are known, perhaps due to indivdually known proper motions or an absolute catalogue position offset relative to the second catalogue, then $h_{\rm pm}$ could simply be a delta function. This would result in the convolution being evaluated with a simple shift in astrometric coordinates, as $(f * \delta)(t) = f(t)$. In practice, however, this is most likely simpler to handle before beginning the cross-match, during the creation of a given catalogue.

4.5.6.2 Extensions to Extra-galactic Source Contamination

In this chapter I have focussed on discussion of the effects of contamination on Galactic sources, focussing on sources with $|b| \leq 10$. These stars suffer much higher average crowding than those sources out of the plane of the Galaxy, and much more crucially need these effects taking into account. However, for catalogues at longer wavelengths

with deep completeness limits, such as *WISE*, faint galaxy source counts will play a role in the perturbation of brighter detections. These perturbations are entirely analogous to the Galactic contamination dealt with in Section 4.3, and extra-galactic sources contribute to the perturbation of sources in the Galactic plane. However, the stellar densities at these Galactic latitudes are much higher than the typical galaxy counts. Additionally, the significant levels of interstellar extinction most significantly affect extra-galactic sources, decreasing their brightnesses more than those of the stars in the Galaxy, further exacerbating the differential source count discrepancy. The contribution of extra-galactic sources to the perturbation of the *WISE* sources considered in this chapter is therefore small. More generally, however, these additional sources from outside the Galaxy can significantly affect the AUFs of these faint, long wavelength catalogues.

This effect is highlighted in Figure 4.18, where the distribution of nearest neighbour matches for Galactic North Pole *Gaia*-*WISE* stars, $b \geq 75$, is shown in black errorbars with $N = 0.014\,\mathrm{mag}^{-1}\mathrm{deg}^{-2}$ (calculated using the differential star counts via equation 4.9), $W1 = 15.5$, and $\sigma_\alpha = 0.11$ arcsecond. For reference a pure Gaussian AUF of the quoted astrometric uncertainty is plotted as the red dotted line. The empirical AUF calculated when taking into account the effects of Galactic *WISE* stars, using the TRILEGAL differential source counts (see Sections 4.3.1 and 4.3.2), is shown as a red dashed line. As can be seen, the Galactic stellar density at the Galactic pole is low (N being some factor of 25 smaller than that typical of the Galactic plane), leading to low astrometric perturbation.

At faint mid-infrared magnitudes, however, the density of galaxies can reach a factor of 10 higher than that of Galactic sources (e.g., figure 7 of Jarrett et al., 2017). Constructing the differential *WISE* galaxy count using the galaxy counts of Jarrett et al. (2017) – see Section 4.3.2.2 for discussion on construction multiple geometric scaling law relationships – the galaxy contaminant empirical AUF is shown in Figure 4.18 as the red solid line. These perturbations produce an AUF in agreement with the distribution of match separations. Therefore, when considering faint, long wavelength detections it is critical that the effects of both Galactic and extra-galactic sources are considered.

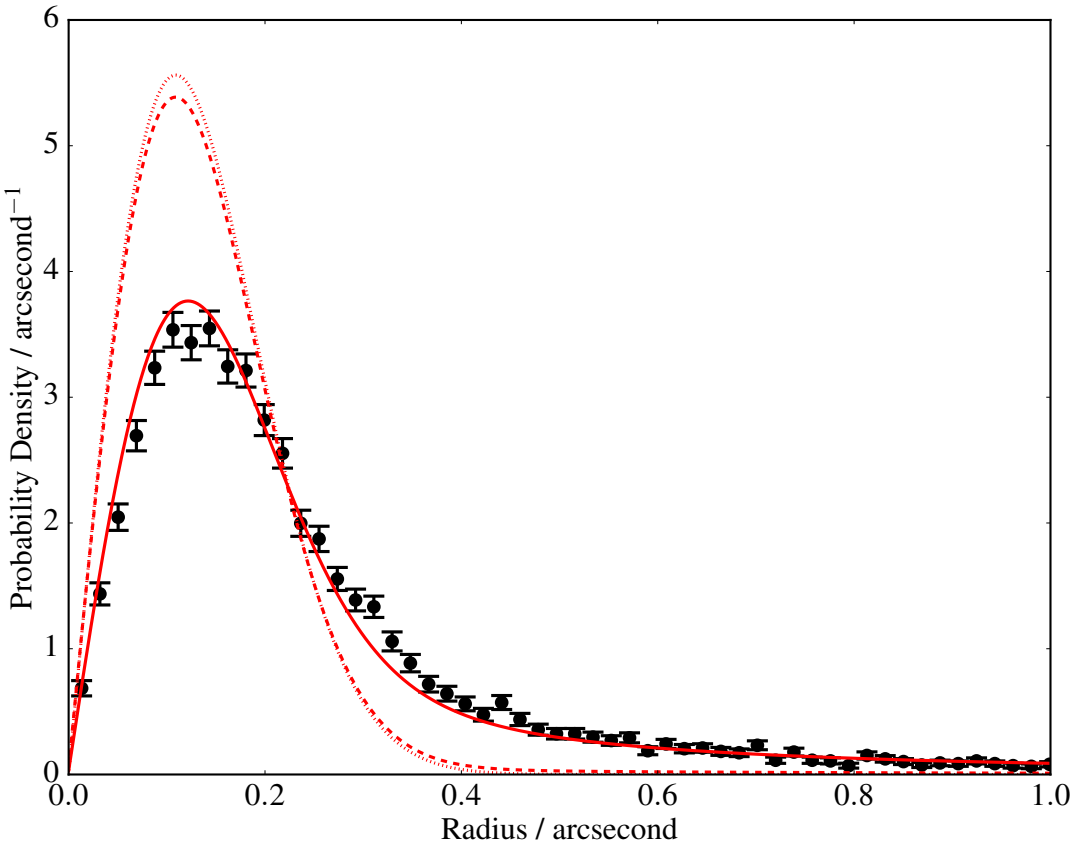


Figure 4.18: *Gaia-WISE* matches for the Galactic North Pole, $b \geq 75$, for $N = 0.009 - 0.019\,\mathrm{mag}^{-1}\mathrm{deg}^{-2}$, $W1 = 15.47 - 15.53$, and $\sigma_\alpha = 0.08 - 0.14$ arcsecond, shown in the black errorbars. The Rayleigh distribution (the representation of a two-dimensional Gaussian in one-dimensional radial coordinates) of the given astrometric uncertainty is shown as a red dotted line. The empirical AUFs including the effects of perturbation from *WISE* Galactic star and galaxy counts are shown as the red dashed and solid lines, respectively. The low density of Galactic sources leads to little perturbation, but the order-of-magnitude higher galaxy counts leads to an AUF in agreement with the distribution of separations.

4.5.7 Further Effects of the Contamination of Stars

I have focussed on the effects the contamination has on the measuring of individual positions throughout this thesis. However, large tails are also seen in the distributions of proper motions (e.g., Dong et al., 2011; Feltzing and Johnson, 2002; Theissen, West, and Dhital, 2016). The large tails seen in contaminated star positions could also propagate to explain the wings of these distributions, as proper motions are simply repeated astrometric measurements over a given time frame. Theissen, West, and Dhital (2016) attempt to account for these contaminant stars, utilising the higher angular resolution of the Sloan Digital Sky Survey (SDSS) data to search for neighbour stars that may be blended in their *WISE* data. However, they only consider SDSS stars on average two magnitudes fainter than the contaminating source, and, as I have shown in this chapter, the contamination of sources at least 10 magnitudes fainter than the central source must be considered for its astrometric perturbations. Any sources which are contaminated by faint background stars will – assuming negligible proper motions of the contaminating sources – have their proper motions reduced by a factor of $1/(1 + F)$, where F is the flux ratio of the contaminants to the central source. This reducing factor is crucial for any proper motion calculations built on catalogues suffering from significant contamination, such as those targeting late-type stars and brown dwarfs (e.g., Schneider et al., 2016). These significantly contaminated objects therefore have much higher proper motions than they appear to exhibit, and this effect should be corrected for. As I showed in this chapter, *WISE* sources in the Galactic plane suffer from an average of 27% contamination, and thus, excluding extremely high proper motion objects which move significant fractions of a PSF image in a reasonable time frame, have proper motions that are 27% higher than those quoted by any *WISE*-based proper motion calculations.

Along much the same lines, this flux contamination reduces the astrometric wobble of sources used to compute parallaxes. The search for nearby brown dwarfs (e.g., Wright et al., 2014) using *NEOWISE* (Mainzer et al., 2014), *WISE*'s reactivation mission, will, in exactly the same way as with its proper motions, calculate a parallax that is perhaps as low

as 78% of its true parallax, overestimating its distance accordingly. Particularly an issue for any survey using *WISE* to calculate source parallaxes, due to its large PSF and faint completeness limit, surveys targeting extremely crowded regions of the Galactic plane will also suffer significant systematics in their proper motion and parallax determination, such as the VISTA Variables in the Via Lactea (VVV) survey of the Galactic bulge (Smith et al., 2018).

The contamination from background sources will also have more subtle implications. While the spectral energy distribution (SED) of the contaminant stars is independent of that of the central source, the stellar mass function is bottom heavy, meaning that most of the contaminants will be of relatively red colours. This will lead to the contamination adding relatively more flux at longer wavelengths than shorter ones. In the case of a relatively early type star, expected to be bright at blue wavelengths and fainter towards the infrared, these contaminants could be confused with other sources of infrared excess. This could lead to cases where source contamination is interpreted as flux from an accretion or protoplanetary disc. The WIRED survey (e.g., Dennihy et al., 2017) therefore expends significant effort following up potential disc excess targets selected with *WISE* to confirm the lack of contaminating flux excesses. This flux contamination could affect the timescales of disc dissipation, targeted direct imaging or interferometry follow-up, or the age of the source due to misinterpretation of signs of youth.

Finally, the flux from contaminating background sources can have implications for the field of exoplanets. The transit of an exoplanet across its host star leads to a calculation of the planet's radius using the ratio of the amount of flux from the star with the planet in front of the star to the star's flux without the planet's shadow. However, if there is additional flux – unrelated to the central star – then the given ratio would be affected. The planet would only pass in front of one of the blended targets, and therefore the transit depth (or change in flux) would be diluted, leading to an interpretation that the planet is smaller than it truly is. This effect is less extreme than in the two previous cases, as the planet radius is related to the square root of the flux, and thus a contaminant contributing 27% additional

flux will result in a planetary radius 89% its true value. Worse still, the inclusion of flux from a star of a different SED could potentially lead to the belief that the blended star is a different spectral type to the true spectral type of the central object, and therefore affect the interpretation of the stellar radius, crucial for converting the flux ratio to a radius ratio. Work has been undertaken on this problem for multiple systems, with Furlan and Howell (2017) considering the influence of *Kepler* companion stars on the densities of detected planets and Cunha et al. (2013) analysing the effect of a fibre-blended companion on the radial velocities of GK stars and thus the masses of orbiting planets. While the influence of multiple star systems on the planetary parameters is large, with Brown (2015) reporting the planet orbiting the visual binary WASP-85 to be twice the naive radius calculated from the transit depth of the equal-magnitude binary system, the effects of background contaminant stars must be taken into account in the calcuation of the radii and masses of extrasolar planets. Work on this in the literature typically involves significant follow-up time, requiring high spatial resolution or detailed spectroscopic analysis (e.g., Southworth and Evans, 2016); however, a significant fraction of the contamination affecting exoplanet parameters could be found significantly quicker and more efficiently with an analysis of the perturbation of the star's AUFs.

4.6 Conclusions

I presented an analysis of the effects of unresolved contaminant stars on the cross-matching of the *Gaia* and *WISE* photometric catalogues. I detailed a treatment of the astrometric uncertainty functions which is capable of folding in these systematic astrometric perturbations in Section 4.3. Comparisons between the ensemble of pairings produced by a probability-based matching process using Gaussian AUFs and the new empirical AUFs were carried out. It was found that without the inclusion of the effects of contamination one in every two *Gaia*-*WISE* matches is rejected. I also detailed the results of a number of test matches, analysing the match rates, false match rates, and effects on the probabilities obtained in Section 4.4.

In addition to discussing the effects these unresolved objects have on the astrometry, in Section 4.5 I considered the effect crowding has on the measured photometric magnitudes. I found that *WISE* objects perturbed sufficiently to be entirely incompatible with a Gaussian AUF are on average 27% brighter than those objects with small astrometric perturbations. Additionally, I compared the *WISE* matches to *Spitzer* detections, using the superior angular resolution of *Spitzer* to resolve the *WISE* contaminants. The ability to resolve the previously blended *WISE* contaminants leads to a skewed intra-*Spitzer* separation distribution. Modelling the effects of hidden contaminants is important for correctly matching two detections which otherwise would have been assumed to be two unphysical individual detections. Moreover, it also allows for the selection of only objects without significant flux from additional sources, critical for comparisons to theoretical models.

Chapter 5

Have We Actually Won the Error Box vs Depth Race?

I've thought of an ending for my book – "And he lived happily ever after... to the end of his days."
— Bilbo Baggins, J. R. R. Tolkien, The Lord of the Rings: The Fellowship of the Ring (1954)

5.1 Applying the Cross-Matching Method in the Future

We are on the precipice of one of the greatest paradigm shifts in the creation of astrophysical catalogues since Galileo. With ever increasingly sensitive surveys, the depth to which we can probe the universe increases much as it ever has these past centuries. However, and crucially, typical error box sizes have stagnated, with a few exceptions, such as that of the *Gaia* mission. The telescopes we are able to build continue to increase in size – continuing the trend of doubling in diameter approximately every 40 years over the past four centuries – and yet, in the case of ground-based observations, are already seeing-limited in resolution. Detector technology has also matured and its positional improvements have already been absorbed into the framework of the scientific observation. We have therefore hit the limit of the ever-increasing ability to pinpoint the position of a source on the sky. The

next generation of survey telescopes will instead offer a significant jump in observation efficiency, with the Large Synoptic Survey Telescope (LSST) offering an almost three orders of magnitude larger field-of-view than the telescopes used by the 2MASS survey, just 20 years on. For the same observation time we will therefore obtain at least an order of magnitude more source detections, but without the corresponding increase in image resolution. Thus, for the first time in history, the "easy" cross-match regime is threatened, and even the optical or near-infrared observations, ever simple to identify the counterparts to, may suffer. Much like the Carte Du Ciel a century ago, the sheer amount of data generated in the next generation of optical and infrared surveys may prove overwhelming, with LSST creating over 20 TB of data *per night*.

It is therefore clear that this new, flexible and robust cross-matching algorithm will become increasingly necessary over the next decade, with LSST a prime example of where its strengths will lie. As I mentioned in Chapter 2, LSST will have a similar crowding Figure of Merit Q to *WISE*, indicating that it will suffer similar crowding to that seen in *WISE* at the faint end of its dynamic range. It will therefore require a cross-match scheme – both internally, across the varying PSF sizes of its passbands, but more importantly externally, when its data are being combined with legacy and supplementary datasets to leverage synergistic wavelength coverage – that can handle the effects of blended sources. In addition, the majority of its sources will be fainter than the *Gaia* completeness limit, and thus lack robust proper motion information, requiring the treatment of separations due to epoch drift on a statistical level, which can be handled by the new formalism for the AUF presented in this thesis, as I discuss in Section 4.5.6.1. The method described here can provide information on the levels of flux contamination, and thus these systematic effects can be removed, allowing for the full potential of this crucial survey to be utilised, without being subject to systematics caused by the crowding of sources by faint contaminant objects. Thus I believe the "full coverage" method described in Section 2.9.2 to be fully applicable to the next generation of faint all-sky surveys.

5.2 Recommendations

Throughout this thesis I have discussed the effects the properties of a photometric catalogue – its dynamic range, resolution, flux calculation methodology, etc. – have on the resulting dataset provided. I summarise here the recommendations made with regards to when various cross-match algorithms are appropriate, considerations that have to be made when using the methods I have described here, and limitations of the algorithms.

1. Consider whether the resulting cross-matched dataset being created is required to not suffer any level of flux contamination. If sources are not permitted to suffer from flux contamination, use a Gaussian as the AUF of the objects (either using $\sigma_{\rm quoted}$ or $\sigma_{\rm core}$, depending on whether the quoted uncertainties of the original catalogues are acceptable), keeping in mind that objects may be flux contaminated but not suffer significant astrometric perturbation. However, if potential flux contamination of sources would not influence the science case for which this cross-match is being performed, can be accounted for, or is simply an unavoidable effect – such as in the case of *WISE* – then use AUFs constructed to include a description of the effects of contaminant stars.

2. Consider the levels to which a given input catalogue is affected by crowding – if both (or all) catalogues are unaffected by crowding to within a given satisfactory level, consider whether the empirical AUF is necessary. *Gaia* is crowded to the tenths of a percent level, IPHAS suffers crowding on the order of a few percent, 2MASS suffers up to 15% crowding, and *WISE* suffers significant crowding even at relatively bright magnitudes, for example. If those few percents of sources astrometrically perturbed to such a level as to introduce additional false negative matches is not an issue, then the simplifying assumption that the AUF is a Gaussian may suffice.

3. The cross-match scheme described in this thesis is only applicable to a two-catalogue match, due in part to the inclusion of counterpart and "field" source densities in the formalism, a function of a specific filter in a specific catalogue, as well as the

inclusion of non-analytic expressions describing the AUFs of the two sources. If a multi-catalogue match is required, first match the two most astrometrically precise datasets, then iteratively merge the latest composite dataset with the next least precise dataset.

4. The creation of the photometric likelihoods using the data rather than comparisons to theoretical models requires good number statistics; the "class" of object must be present in both catalogues to some minimum threshold to populate the magnitude-magnitude parameter space with sufficient precision. If the cross-matching being undertaken is to search for rare or atypical sources, exhibiting highly unusual spectral energy distributions, consider whether the creation of theoretical models to describe the class of object would be more applicable.

5. The current implementation of this cross-match scheme makes the assumption that the PSF and corresponding covariance matrix of a detection are circular; if one of the catalogues being merged suffers from significant off-axis effects due to, e.g., telescope optics, this assumption may not hold. *WISE*, for example, suffers <10% off-axis asymmetric in >90% of detections.

6. The creation of an empirical AUF requires a description of the magnitude distribution of potential contaminating sources. In the Galactic plane this could be done using Galaxy simulations, but at the Galactic poles a description of galaxy counts should be used. Consideration must be given to all potential sources of contaminants – mid-way out of the plane of the Galaxy both Galactic stars and extra-galactic sources are likely to exist in roughly equal numbers.

7. The level to which catalogues being cross-matched can potentially be deblended must be considered when merging datasets. For example, the *WISE* pipeline is restricted to the deblending of up to one additional component, up to 2.5 magnitudes fainter than the primary source, which places an upper limit on how much the catalogue is able to overcome the effects of crowding. A more rigorous pipeline,

allowing further component deblending, would be able to alleviate the effects of crowding more than one which does not allow source deblending at all.

8. The method described in this thesis requires that the differential source counts at the bright end of a given catalogue is described by a geometric scaling law with scaling parameter $z = 2$. In regions of extreme crowding, the significant flux brightening of sources due to the blending of many contaminating objects leads to a differential source counts that does not follow this law, and the cross-match scheme described here will produce matches that cannot necessarily be trusted.

9. When using the "full coverage" method the average flux contamination of the empirical AUFs used in the matching process is obtained. This information can be used to remove sources contaminated above a critical threshold; consideration must therefore be given to what that critical threshold is.

10. If proper motions are likely to be an issue in the cross-match under consideration – because high proper motion objects are the main focus, or if the precision of the datasets means that even relatively low proper motions are comparable to the length scales of the matches – then they must be handled. If, in the case of *Gaia*, individual proper motions are known, they can be applied on a source-by-source basis, or if they are not known, a statistical distribution of the offsets due to proper motions by an ensemble of objects (from, e.g., simulations) can be folded into the empirical AUF in the new formalism described in this thesis.

11. If both catalogues being cross-matched suffer from extreme crowding, the number of sources in a given "island" may become computationally unmanageable. If this occurs, an upper limit should be placed on the island permutation, and the over-sized islands either removed entirely or split at their largest inter-catalogue separation, creating two smaller sub-islands.

12. The additional flux introduced by contaminating sources will affect secondary parameters derived for the central sources, dampening the effects due to proper motion,

parallax, and exoplanet transit measurements. If the catalogues in question will go on to be used to derive any of these parameters, consider the contamination levels of the chosen catalogues. If the level of contamination of the sources in question is too high, consider using a different, higher angular resolution dataset; otherwise the effects can be remedied somewhat by accounting for the flux contamination present in the empirical AUFs used in the cross-matching process.

5.3 Technical Implementation

As discussed in Section 1.5, the catalogue cross-matching method laid out in this thesis is concerned with the extreme cases of sky crowding, in which false match rates (both positive and negative) would be at their highest. These are overcome with the two-directional implementation of an astrophysical-model-free creation of photometric likelihoods, and a new, robust description of the AUF, allowing for the inclusion of systematic causes of sky separation – the most likely, in these crowded fields, being the centroid shifts caused by blended sources. However, these improvements come at a cost. The methods described here are very complex – both mathematically and computationally – when compared to other cross-matching methods, the cost of removing the caveats and assumptions (such as one catalogue being sparse enough for all of its detections to be considered independent) made by these simpler algorithms. The creation of empirical AUFs including the effects of perturbation due to faint contaminant stars requires numerical modelling, and thus the AUF convolution, required to calculate the probability of two detections having some separation given the hypothesis that they are two detections of one physical object, must also be numerical. This results in significant computational cost compared to the trivial analytical evaluation of the convolution of two Gaussians. Additionally, the "in situ" derivation of the photometric likelihoods requires significant run time compared to the theoretical modelling required by the methodology laid out by Budavári and Szalay (2008).

The runtimes for the matches used in Chapters 3 and 4 – ranging from approximately 25 to 50 square degrees – took somewhere from an hour to 90 minutes. They were run

on 16 Intel Xeon E5-2697 CPUs clocked at 2.70GHz, hyper-threaded to 32 threads, with 128GB of RAM, although in practice these small runs, matching on the order of a million stars in each catalogue, only required at most 10GB of RAM. The extrapolation to an all-sky catalogue cross-match, for the 1.1 billion *Gaia* DR1 sources and 750 million *WISE* detections is on the order of 2-3 weeks, and thus requires some level of NumPy memory-mapping to ensure all arrays can be maintained in memory were necessary. Thus the trade-off for ensuring the most robust and reliable results, even in the most confused regions of the sky, is computational time (for comparison, Marrese et al., 2017 state their all-sky *Gaia* DR1-*WISE* cross-match, using a simple, purely astrometric likelihood ratio method, took a mere 7.5 hours). The code was written in Python 2.7, utilising various Python module functions (NumPy, SciPy, astropy) for array handling, astrometric coordinate manipulation, and mathematical functions. Due to the complexity of the numerical integration, and increase potential counterpart numbers due to the large non-Gaussian wings to the AUFs, significant parts of the computations are passed out to Fortran via f2py, using OPENMP to parallelise the numerical modelling of the PSF centroid shifts, numerical convolutions, and creation of photometric likelihoods b and f. Raw catalogues were cleaned and processed in Python, creating NumPy arrays to store a subset of the provided information – astrometric positions, magnitudes, and detection quality flags (see Table 1.2).

While these computational costs are higher than those of comparable cross-match algorithms, they are small compared with the overall costs of a large survey program or telescope mission. The additional time may in fact be well invested to ensure robust matches, applicable to a wide range of scientific research interests, even in the most crowded of regions. As surveys probe increasingly faint populations, increasing crowding to extreme levels for even reasonably high angular resolution telescopes, the methodology laid out here may well become necessary, with the extra few weeks spent applying a cross-matching algorithm negligible compared with the years spent conducting a survey – LSST, for example, will release a data product yearly through its 10-year planned operation.

5.4 Final Remarks

To overcome the upheaval in the error box vs depth race and continue to ever fainter magnitudes, we must turn to the "out of the box" thinking developed in application to the more troublesome depths of the cross-matching history. I have presented here a new and novel approach to the consideration of the cross-matching of two photometric catalogues. The use of the photometric information in both photometric catalogues will allow for improved differentiation between the multiple potential counterparts each source will have to consider as its source identification in the opposing catalogue. Additionally, I have created a formalism for the treatment of the perturbation of bright sources by faint contaminants, crucial as the depths of catalogues created from incredibly sensitive surveys reach extremely crowded levels, even for otherwise relatively high angular resolution telescopes. *WISE* is highlighted as a particularly extreme case of crowding for the current generation of telescopes, with a large point-spread function and faint completeness limit; and yet this is the typical crowding of the future telescope, with LSST suffering similar levels of crowding at its faint magnitude limit as *WISE* does at its.

The work outlined here is therefore crucial to the robust matching of sources between catalogues, both of the next generation and legacy surveys. The astrometric perturbations suffered can reveal the extent to which sources are photometrically compromised, and allow for these additional sources of flux to be accounted for, removing them from systematically affecting any astrophysical parameter derived from the detected brightness of a source. We now have a framework with which to recover a significant fraction of matches that would otherwise be lost by a more naive cross-match, and can actually use this information to gain further insight into the nature of the observation and correct for its systematics.

It seems that we have not won the error box vs depth race quite so conclusively as it appeared at the dawn of the new millennium. Space is crowded, and yet hitherto has not been considered thus for the most part, with rare consideration in the literature. However, building on centuries of progress, both technological and mathematical, we can change the rules of the competition.

Bibliography

Airy, G. B. (1835). “On the Diffraction of an Object-glass with Circular Aperture”. *Transactions of the Cambridge Philosophical Society*, 5, 283.

Aschenbach, B. et al. (1981). “The ROSAT mission”. Space Sci. Rev. 30, 569.

Bahcall, J. N. and R. M. Soneira (1980). “Star counts as an indicator of galactic structure and quasar evolution”. ApJL, 238, L17.

Barentsen, G. et al. (2014). “The second data release of the INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS DR2)”. MNRAS, 444, 3230.

Bayes, Thomas (1763). “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S”. *Philosophical Transactions*, 53, 370.

Bessel, F. W. (1844). “On the variations of the proper motions of Procyon and Sirius”. MNRAS, 6, 136.

— (1848). *Populare Vorlesungen uber Wissenschaftliche Gegenstande*.

Bessel, F. W. and J. Bradley (1818). *Fundamenta astronomiae pro anno MDCCLV : deducta ex observationibus per annos 1750-1762 institutis viri incomparabilis James Bradley in specula astronomica Grenovicensi*.

Bourne, N. et al. (2016). “The Herschel-ATLAS Data Release 1 - II. Multi-wavelength counterparts to submillimetre sources”. MNRAS, 462, 1714.

Bouvier, J. and I. Appenzeller (1992). “A magnitude-limited spectroscopic and photometric survey of Rho Ophiuchus X-ray sources”. A&AS, 92, 481.

Boyle, W. S. and G. E. Smith (1970). “Charge Coupled Semiconductor Devices”. *Bell System Technical Journal*, 49, 4, 587.

Bradley, James (1798). *Observations 1750 - 1762 (Volume 1. Bradley).*

Bradley, James and Nathaniel Bliss (1805). *Observations 1750 - 1764 (Volume 2. Bradley & Bliss).*

Brahe, T. (1602). *Tychonis Brahe Astronomiæ instauratæ mechanica.*

Broos, P. S., K. V. Getman, M. S. Povich, E. D. Feigelson, L. K. Townsley, T. Naylor, M. A. Kuhn, R. R. King, and H. A. Busk (2013). "Identifying Young Stars in Massive Star-forming Regions for the MYStIX Project". ApJS, 209, 32.

Brown, D. J. A. (2015). "Discovery of WASP-85 Ab: A Hot Jupiter in a Visual Binary System". *European Planetary Science Congress*, 10, 603.

Brusa, M. et al. (2005). "XMM-Newton observations of Extremely Red Objects and the link with luminous, X-ray obscured quasars". A&A, 432, 69.

Brusa, M. et al. (2007). "The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources". ApJS, 172, 353.

Buckley, D. A. H., I. R. Tuohy, and R. A. Remillard (1985). "Optical counterparts of HEAO-1 X-ray sources". *Proceedings of the Astronomical Society of Australia*, 6, 147.

Budavári, T. and A. Basu (2016). "Probabilistic Cross-identification in Crowded Fields as an Assignment Problem". AJ, 152, 86.

Budavári, T. and T. J. Loredo (2015). "Probabilistic Record Linkage in Astronomy: Directional Cross-Identification and Beyond". *Annual Review of Statistics and Its Application*, 2, 113.

Budavári, T. and A. S. Szalay (2008). "Probabilistic Cross-Identification of Astronomical Sources". ApJ, 679, 301.

Chang, C.-K., C.-M. Ko, and T.-H. Peng (2010). "The Information of the Milky Way From Two Micron All Sky Survey Whole Sky Star Count: The Luminosity Function". ApJ, 724, 182.

Cowie, L. L., J. P. Gardner, S. J. Lilly, and I. McLean (1990). "A K band deep galaxy survey". ApJL, 360, L1.

Cowley, A. P., P. C. Schmidtke, A. L. Anderson, and T. K. McGrath (1995). "Determination of the optical counterpart of LMC X-1". PASP, 107, 145.

Cunha, D., P. Figueira, N. C. Santos, C. Lovis, and G. Boué (2013). "Impact of stellar companions on precise radial velocities". A&A, 550, A75.

Curto, A., M. Tucci, J. González-Nuevo, L. Toffolatti, E. Martínez-González, F. Argüeso, A. Lapi, and M. López-Caniego (2013). "Forecasts on the contamination induced by unresolved point sources in primordial non-Gaussianity beyond Planck". MNRAS, 432, 728.

Cutri, R. M. et al. (2012). *Explanatory Supplement to the WISE All-Sky Data Release Products*. Tech. rep.

Danziger, I. J. et al. (1990). "Optical identification content of the ROSAT all sky survey." *The Messenger*, 62, 4.

Daylan, T., S. K. N. Portillo, and D. P. Finkbeiner (2017). "Inference of Unresolved Point Sources at High Galactic Latitudes Using Probabilistic Catalogs". ApJ, 839, 4.

de Moivre, A. (1718). *Doctrine of Chances: Or, A Method of Calculating the Probability of Events in Play*.

— (1733). *Approximatio ad summam terminorum binomii [mathematical expression] in seriem expansi*.

de Ruiter, H. R., A. G. Willis, and H. C. Arp (1977). "A Westerbork 1415 MHz survey of background radio sources. II - Optical identifications with deep IIIA-J plates". A&AS, 28, 211.

Debes, J. H., D. W. Hoard, S. Wachter, D. T. Leisawitz, and M. Cohen (2011). "The WIRED Survey. II. Infrared Excesses in the SDSS DR7 White Dwarf Catalog". ApJS, 197, 38.

Dennihy, E., J. C. Clemens, J. H. Debes, B. H. Dunlap, D. Kilkenny, P. C. O'Brien, and J. T. Fuchs (2017). "WIRED for EC: New White Dwarfs with WISE Infrared Excesses and New Classification Schemes from the Edinburgh-Cape Blue Object Survey". ApJ, 849, 77.

Dewhirst, D. W. (1959). "The optical identification of radio sources". In: *URSI Symp. 1: Paris Symposium on Radio Astronomy*. Ed. by R. N. Bracewell. Vol. 9. IAU Symposium, 507.

Dick, W. R., H.-J. Tucholke, P. Brosche, R. Galas, M. Geffert, and J. Guibert (1993). "HIPPARCOS link with Carte du Ciel triple images". A&A, 279, 267.

Dong, R., J. Gunn, G. Knapp, C. Rockosi, and M. Blanton (2011). "Investigation of the Errors in Sloan Digital Sky Survey Proper-motion Measurements Using Samples of Quasars". AJ, 142, 116.

Doyle, M. T. et al. (2005). "The HIPASS catalogue - III. Optical counterparts and isolated dark galaxies". MNRAS, 361, 34.

Drew, J. E. et al. (2005). "The INT Photometric Hα Survey of the Northern Galactic Plane (IPHAS)". MNRAS, 362, 753.

Edge, D. O., J. R. Shakeshaft, W. B. McAdam, J. E. Baldwin, and S. Archer (1959). "A survey of radio sources at a frequency of 159 Mc/s." MmRAS, 68, 37.

ESA, ed. (1997). *The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission*. Vol. 1200. ESA Special Publication.

Fazio, G. G. et al. (2004). "The Infrared Array Camera (IRAC) for the Spitzer Space Telescope". ApJS, 154, 10.

Feigelson, E. D. and G. A. Kriss (1989). "Soft X-ray observations of pre-main-sequence stars in the Chamaeleon dark cloud". ApJ, 338, 262.

Feltzing, S. and R. A. Johnson (2002). "A new, cleaner colour-magnitude diagram for the metal-rich globular cluster NGC 6528. Velocity dispersion in the Bulge, age and proper motion of NGC 6528". A&A, 385, 67.

Flamsteed, J. (1725). *Historia Coelestis Britannica: Complectens Praefationem spatiosam (sive in Stellarum Fixarum Catalogum Prolegomena) quae brevem Astronomiae Historiam praebet, atque Descriptionem Observationum peractarum, & Organorum adhibitorum tum a pristinis Astronomis, tum In Observatorio Regio Grenovicensi ; Dein-*

ceps Fixarum Catalogum a Ptolemaeo, Uleg Beig, Tychone Brahaeo, Gulielmo Hessiae Landtgravio, ac Hevelio constructum ... : Quibus adnexus est Fixarum quarundam Australium Catalogus, in nostro Hemisphaerio non adspectabilium, Denique Tabulae .... 3. Meere.

Flesch, E. and M. J. Hardcastle (2004). "An all-sky optical catalogue of radio/X-ray sources". A&A, 427, 387.

Fleuren, S. et al. (2012). "Herschel-ATLAS: VISTA VIKING near-infrared counterparts in the Phase 1 GAMA 9-h data". MNRAS, 423, 2407.

Forrest, W. J., A. Moneti, C. E. Woodward, J. L. Pipher, and A. Hoffman (1985). "The New Near Infrared Array Camera at the University of Rochester". PASP, 97, 183.

Fotopoulou, S. et al. (2016). "The XXL Survey. VI. The 1000 brightest X-ray point sources". A&A, 592, A5.

Furlan, E. and S. B. Howell (2017). "The Densities of Planets in Multiple Stellar Systems". AJ, 154, 66.

Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. de Bruijne, F. Mignard, R. Drimmel, and 5. co-authors (2016a). "Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties". A&A, 595, 1, 2.

Gaia Collaboration et al. (2016b). "The Gaia mission". A&A, 595, A1.

Galilei, G. (1610). *Sidereus nuncius magna, longeque admirabilia spectacula pandens lunae facie, fixis innumeris, lacteo circulo, stellis nebulosis, ... Galileo Galileo : nuper a se reperti beneficio sunt observata in apprime vero in quatuor planetis circa Iovis stellam disparibus intervallis, atque periodis, celeritate mirabili circumvolutis ... atque Medicea sidera nuncupandos decrevit*.

Garcia, M. R., J. E. Grindlay, C. D. Bailyn, J. L. Pipher, M. A. Shure, and C. E. Woodward (1992). "The infrared counterpart of GX 13 + 1". AJ, 103, 1325.

Gauss, K. F. (1809). *Theoria motvs corporvm coelestivm in sectionibvs conicis solem ambientivm.*

Gavras, P., D. Sinachopoulos, J. F. Le Campion, and C. Ducourant (2010). "The CPMDS catalogue of common proper motion double stars in the Bordeaux Carte du Ciel zone". A&A, 521, A4.

Geffert, M., P. Bonnefond, G. Maintz, and J. Guibert (1996). "The astrometric accuracy of "Carte du Ciel" plates and proper motions in the field of the open cluster NGC 1647." A&AS, 118, 277.

Giacconi, R., H. Gursky, and J. R. Waters (1964). "Two Sources of Cosmic X-rays in Scorpius and Sagittarius". Nature, 204, 981.

Giacconi, R., H. Gursky, F. R. Paolini, and B. B. Rossi (1962). "Evidence for x Rays From Sources Outside the Solar System". *Physical Review Letters*, 9, 439.

Girardi, L., M. A. T. Groenewegen, E. Hatziminaoglou, and L. da Costa (2005). "Star counts in the Galaxy. Simulating from very deep to very shallow photometric surveys with the TRILEGAL code". A&A, 436, 895.

Gray, A., G. B. Mathews, and T. M. Macrobert (1895). *A Treatise on Bessel Functions and Their Applications to Physics*.

Gursky, H., R. Giacconi, F. R. Paolini, and B. B. Rossi (1963). "Further Evidence for the Existence of Galactic x Rays". *Physical Review Letters*, 11, 530.

Haakonsen, C. B. and R. E. Rutledge (2009). "XID II: Statistical Cross-Association of ROSAT Bright Source Catalog X-ray Sources with 2MASS Point Source Catalog Near-Infrared Sources". ApJS, 184, 138.

Halley, E. (1717). "Considerations on the Change of the Latitudes of Some of the Principal Fixt Stars. By Edmund Halley, R. S. Sec." *Philosophical Transactions of the Royal Society of London Series I*, 30, 736.

Hazard, C., M. B. Mackey, and A. J. Shimmins (1963). "Investigation of the Radio Source 3C 273 By The Method of Lunar Occultations". Nature, 197, 1037.

Henden, A. and U. Munari (2014). "The APASS all-sky, multi-epoch BVgri photometric survey". *Contributions of the Astronomical Observatory Skalnate Pleso*, 43, 518.

Herschel, J. F. W. (1857). *Essays from the Edinburgh and Quarterly Reviews, with Addresses and Other Pieces*.

Hey, J. S. (1946). "Solar Radiations in the 4-6 Metre Radio Wave-Length Band". Nature, 157, 47.

Heyer, M. H., C. Brunt, R. L. Snell, J. E. Howe, F. P. Schloerb, and J. M. Carpenter (1998). "The Five College Radio Astronomy Observatory CO Survey of the Outer Galaxy". ApJS, 115, 241.

Hogg, D. W. (2001). "Confusion Errors in Astrometry and Counterpart Association". AJ, 121, 1207.

Hunter, John D. (2007). "Matplotlib: A 2D Graphics Environment". *Computing in Science & Engineering*, 9, 90.

Jansen, F. et al. (2001). "XMM-Newton observatory. I. The spacecraft and operations". A&A, 365, L1.

Jansky, K. G. (1933). "Radio Waves from Outside the Solar System". Nature, 132, 66.

Jarrett, T. H. et al. (2017). "Galaxy and Mass Assembly (GAMA): Exploring the WISE Web in G12". ApJ, 836, 182.

Jaynes, E. T. and G. L. Bretthorst (2003). *Probability Theory*.

Jeong, W.-S., C. P. Pearson, H. M. Lee, S. Pak, and T. Nakagawa (2006). "Far-infrared detection limits - II. Probing confusion including source confusion". MNRAS, 369, 281.

Johnson, H. L. (1962). "Infrared Stellar Photometry." ApJ, 135, 69.

Jones, E., E. Oliphant, P Peterson, et al. (2001). *SciPy: Open Source Scientific Tools for Python*.

Kellogg, K., S. Metchev, K. Geißler, S. Hicks, J. D. Kirkpatrick, and R. Kurtev (2015). "A Targeted Search for Peculiarly Red L and T Dwarfs in SDSS, 2MASS, and WISE: Discovery of a Possible L7 Member of the TW Hydrae Association". AJ, 150, 182.

Kepler, J. and T Brahe (1627). *Tabvlae Rudolphinae, qvibvs astronomicae scientiae, temporum longinquitate collapsae restauratio continetur*.

Kerekes, G., T. Budavári, I. Csabai, A. J. Connolly, and A. S. Szalay (2010). "Cross Identification of Stars with Unknown Proper Motions". ApJ, 719, 59.

Kessler, M. F. et al. (1996). "The Infrared Space Observatory (ISO) mission." A&A, 315, L27.

King, I. R. (1983). "Accuracy of measurement of star images on a pixel array". PASP, 95, 163.

Krawczyk, C. M., G. T. Richards, S. S. Mehta, M. S. Vogeley, S. C. Gallagher, K. M. Leighly, N. P. Ross, and D. P. Schneider (2013). "Mean Spectral Energy Distributions and Bolometric Corrections for Luminous Quasars". ApJS, 206, 4.

Kukarkin, B. V., Y. N. Efremov, M. S. Frolov, G. I. Medvedeva, P. N. Kholopov, N. E. Kurochkin, N. P. Kukarkina, N. B. Perova, and V. P. Fedorovich (1968). "Identification List of the New Variable Stars Nominated in 1968". *Information Bulletin on Variable Stars*, 311.

Kunkel, W., P. Osmer, M. Smith, A. Hoag, D. Schroeder, W. A. Hiltner, H. Bradt, S. Rappaport, and H. W. Schnopper (1970). "An Optical Search for the X-Ray Sources GX3+1, GX5-1, GX9+1, and GX17+2". ApJL, 161, L169.

Laplace, Pierre-Simon (1774). *Mémoire sur la probabilité des causes par les évenments*.

— (1820). *Théorie analytique des probabilités*.

Leggett, S. K. and M. R. S. Hawkins (1989). "Low mass stars in the region of the Hyades cluster". MNRAS, 238, 145.

Lindegren, L. et al. (2016). "Gaia Data Release 1. Astrometry: one billion positions, two million proper motions and parallaxes". A&A, 595, A4.

Line, J. L. B., R. L. Webster, B. Pindor, D. A. Mitchell, and C. M. Trott (2017). "PUMA: The Positional Update and Matching Algorithm". *PASA*, 34, 3.

López, K. M., M. Heida, P. G. Jonker, M. A. P. Torres, T. P. Roberts, D. J. Walton, D.-S. Moon, and F. A. Harrison (2017). "A systematic search for near-infrared counterparts of nearby ultraluminous X-ray sources (II)". MNRAS, 469, 671.

Lynch, B. M. and L. H. Robbins (1978). "Namoratunga: The First Archeoastronomical Evidence in Sub-Saharan Africa". *Science*, 200, 766.

Mainzer, A. et al. (2014). "Initial Performance of the NEOWISE Reactivation Mission". ApJ, 792, 30.

Malkov, O. and S. Karpov (2011). "Cross-Matching Large Photometric Catalogs for Parameterization of Single and Binary Stars". In: *Astronomical Data Analysis Software and Systems XX*. Ed. by I. N. Evans, A. Accomazzi, D. J. Mink, and A. H. Rots. Vol. 442. Astronomical Society of the Pacific Conference Series, 583.

Mann, R. G. et al. (1997). "Observations of the Hubble Deep Field with the Infrared Space Observatory - IV. Association of sources with Hubble Deep Field galaxies". MNRAS, 289, 482.

Marquez, M. J., T. Budavári, and L. M. Sarro (2014). "Improving cross-identification of galaxies using their photometry". A&A, 563, A14.

Marrese, P. M., S. Marinoni, M. Fabrizio, and G. Giuffrida (2017). "Gaia Data Release 1. Cross-match with external catalogues. Algorithm and results". A&A, 607, A105.

Mason, K. O. et al. (1995). "Optical identification of EUV sources from the ROSAT Wide Field Camera all-sky survey". MNRAS, 274, 1194.

Matthews, T. A. and A. R. Sandage (1963). "Optical Identification of 3C 48, 3C 196, and 3C 286 with Stellar Objects." ApJ, 138, 30.

Michalik, D., L. Lindegren, and D. Hobbs (2015). "The Tycho-Gaia astrometric solution. How to get 2.5 million parallaxes with less than one year of Gaia data". A&A, 574, A115.

Miller, B. W., B. Margon, and M. G. Burton (1993). "The infrared counterpart of the bright X-ray source GX340+0". AJ, 106, 28.

Mocanu, L. M. et al. (2013). "Extragalactic Millimeter-wave Point-source Catalog, Number Counts and Statistics from 771 $\deg^2$ of the SPT-SZ Survey". ApJ, 779, 61.

Morales, E. F. E. and T. P. Robitaille (2017). "Do individual Spitzer young stellar object candidates enclose multiple UKIDSS sources?" A&A, 598, A136.

Morrison, L. V. and P. Gibbs (1986). "Carlsberg Automatic Transit Circle - First Two Test Catalogues and the Programme for La-Palma". In: *Astrometric Techniques*. Ed. by H. K. Eichhorn and R. J. Leacock. Vol. 109. IAU Symposium, 497.

Munari, U., P. Ochner, A. Siviero, M. Graziani, S. Dallaporta, G. L. Righetti, G. Cherini, and F. Castellani (2013). "Nova Cephei 2013 has emerged from dust obscuration". *The Astronomer's Telegram*, 5389.

Munari, U. et al. (2014). "APASS Landolt-Sloan BVgri Photometry of RAVE Stars. I. Data, Effective Temperatures, and Reddenings". AJ, 148, 81.

Murakami, H. et al. (2007). "The Infrared Astronomical Mission AKARI*". PASJ, 59, S369.

Naylor, T. (1998). "An optimal extraction algorithm for imaging photometry". MNRAS, 296, 339.

Naylor, T., P. S. Broos, and E. D. Feigelson (2013). "Bayesian Matching for X-Ray and Infrared Sources in the MYStIX Project". ApJS, 209, 30.

Naylor, T., P. A. Charles, and A. J. Longmore (1991). "Infrared observations of low-mass X-ray binaries. I - Candidates for bright bulge sources". MNRAS, 252, 203.

Neugebauer, G. et al. (1984). "The Infrared Astronomical Satellite (IRAS) mission". ApJL, 278, L1.

Ogle, P. M., J. Mazzarella, R. Ebert, D. Fadda, T. Lo, S. Terek, M. Schmitz, and NED Team (2015). "Rule-based Cross-matching of Very Large Catalogs". In: *Astronomical Data Analysis Software and Systems XXIV (ADASS XXIV)*. Ed. by A. R. Taylor and E. Rosolowsky. Vol. 495. Astronomical Society of the Pacific Conference Series, 25.

Olsen, K. A. G., R. D. Blum, and F. Rigaut (2003). "Stellar Crowding and the Science Case for Extremely Large Telescopes". AJ, 126, 452.

Peterson, B. A. (1972). "A search for the optical counterpart of Centaurus X-3". *Proceedings of the Astronomical Society of Australia*, 2, 110.

Peterson, Pearu (2009). "F2PY: a tool for connecting Fortran and Python programs". *International Journal of Computational Science and Engineering*, 4, 296.

Pineau, F.-X. et al. (2017). "Probabilistic multi-catalogue positional cross-match". A&A, 597, A89.

Plewa, P M and R Sari (2018). "Unrecognized astrometric confusion in the Galactic Centre". MNRAS, 476, 4, 4372.

Ptolemaeus, Claudius (1515). *Almagestum*.

Rayleigh, L. (1880). "Investigations in optics, with special reference to the spectroscope". MNRAS, 40, 254.

Reber, G. (1944). "Cosmic Static." ApJ, 100, 279.

Richter, G. A. (1975). "Search for optical identifications in the 5C3-radio survey. II - Statistical treatment and results". *Astronomische Nachrichten*, 296, 65.

Ross, F. E. (1922). "The Relation Between the Diameter of A Photographic Star Image and its Magnitude". In: *Publications of the American Astronomical Society*. Vol. 4. Publications of the American Astronomical Society, 280.

Rutledge, R. E., R. J. Brunner, T. A. Prince, and C. Lonsdale (2000). "XID: Cross-Association of ROSAT/Bright Source Catalog X-Ray Sources with USNO A-2 Optical Point Sources". ApJS, 131, 335.

Salvato, M., J. Buchner, T. Budavári, T. Dwelly, A. Merloni, M. Brusa, A. Rau, S. Fotopoulou, and K. Nandra (2018). "Finding counterparts for all-sky X-ray surveys with NWAY: a Bayesian algorithm for cross-matching multiple catalogues". MNRAS, 473, 4937.

Sandage, A. et al. (1966). "On the optical identification of SCO X-1". ApJ, 146, 316.

Sato, Y., L. L. Cowie, K. Kawara, Y. Taniguchi, Y. Sofue, H. Matsuhara, and H. Okuda (2002). "Mid-Infrared Identification of Faint Submillimeter Sources". ApJL, 578, L23.

Schmidt, M. (1963). "3C 273 : A Star-Like Object with Large Red-Shift". Nature, 197, 1040.

Schneider, A. C., J. Greco, M. C. Cushing, J. D. Kirkpatrick, A. Mainzer, C. R. Gelino, S. B. Fajardo-Acosta, and J. Bauer (2016). "A Proper Motion Survey Using the First Sky Pass of NEOWISE-reactivation Data". ApJ, 817, 112.

Seares, F. H. (1914). “Relation of the Mount Wilson Photographic and Photo-Visual Magnitude Scales”. PASP, 26, 211.

Shakeshaft, J. R., M. Ryle, J. E. Baldwin, B. Elsmore, and J. H. Thomson (1955). “A survey of radio sources between declinations -38 and +83.” MmRAS, 67, 106.

Shklovsky, I. S. (1967). “On the Nature of the Source of X-Ray Emission of Sco XR-1.” ApJL, 148, L1.

Skrutskie, M. F. et al. (2006). “The Two Micron All Sky Survey (2MASS)”. AJ, 131, 1163.

Smith, B. A. (1976). “Astronomical imaging applications for CCDs”. In: *Charge-Coupled Device Technology and Applications*. Ed. by S. Iwasa and W. J. White.

Smith, L. C. et al. (2018). “VIRAC: the VVV Infrared Astrometric Catalogue”. MNRAS, 474, 1826.

Southworth, G.C. (1945). “Microwave radiation from the sun”. *Journal of the Franklin Institute*, 239, 4, 285.

Southworth, J. and D. F. Evans (2016). “Contamination from a nearby star cannot explain the anomalous transmission spectrum of the ultrashort period giant planet WASP-103 b”. MNRAS, 463, 37.

Steeghs, D. and J. Casares (2002). “The Mass Donor of Scorpius X-1 Revealed”. ApJ, 568, 273.

Stocke, J. T., J. Liebert, I. M. Gioia, T. Maccacaro, R. E. Griffiths, I. J. Danziger, D. Kunth, and J. Lub (1983). “The Einstein Observatory Medium Sensitivity Survey - Optical identifications for a complete sample of X-ray sources”. ApJ, 273, 458.

Sutherland, W. and W. Saunders (1992). “On the likelihood ratio for source identification”. MNRAS, 259, 413.

Taylor, B. G., R. D. Andresen, A. Peacock, and R. Zobl (1982). “The European X-ray observatory Exosat - Its mission and scientific instruments”. *ESA Bulletin*, 31, 20.

Theissen, C. A., A. A. West, and S. Dhital (2016). “Motion Verified Red Stars (MoVeRS): A Catalog of Proper Motion Selected Low-mass Stars from WISE, SDSS, and 2MASS”. AJ, 151, 41.

Turner, H. H. (1912). *The Great Star Map: Being a General Account of the International Project Known as the Astrographic Chart*, 159.

van Altena, W. F. (2012). *Astrometry for Astrophysics*.

van der Walt, Stéfan, S. Chris Colbert, and Gaël Varoquaux (2011). "The NumPy Array: A Structure for Efficient Numerical Computation". *Computing in Science & Engineering*, 13, 22.

Webb, T. M. A., S. J. Lilly, D. L. Clements, S. Eales, M. Yun, M. Brodwin, L. Dunne, and W. K. Gear (2003). "The Canada-UK Deep Submillimeter Survey. VII. Optical and Near-Infrared Identifications for the 14 Hour Field". ApJ, 597, 680.

Webster, B. L., W. L. Martin, M. W. Feast, and P. J. Andrews (1972). "Optical Candidate for SMC X-1". *Nature Physical Science*, 240, 183.

Weisskopf, M. C., B. Brinkman, C. Canizares, G. Garmire, S. Murray, and L. P. Van Speybroeck (2002). "An Overview of the Performance and Scientific Results from the Chandra X-Ray Observatory". PASP, 114, 1.

Werner, M. W. et al. (2004). "The Spitzer Space Telescope Mission". ApJS, 154, 1.

Wolstencroft, R. D., A. Savage, R. G. Clowes, H. T. MacGillivray, S. K. Leggett, and M. Kalafi (1986). "The Identification of IRAS Point Sources - Part One - a 304-DEGREE Field Centred on the South Galactic Pole". MNRAS, 223, 279.

Wright, E. L. et al. (2010). "The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance". AJ, 140, 1868.

Wright, E. L. et al. (2014). "NEOWISE-R Observation of the Coolest Known Brown Dwarf". AJ, 148, 82.